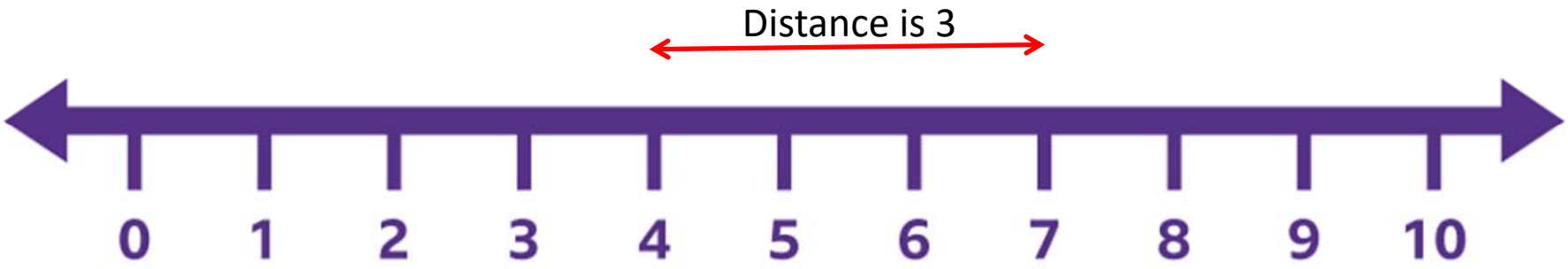


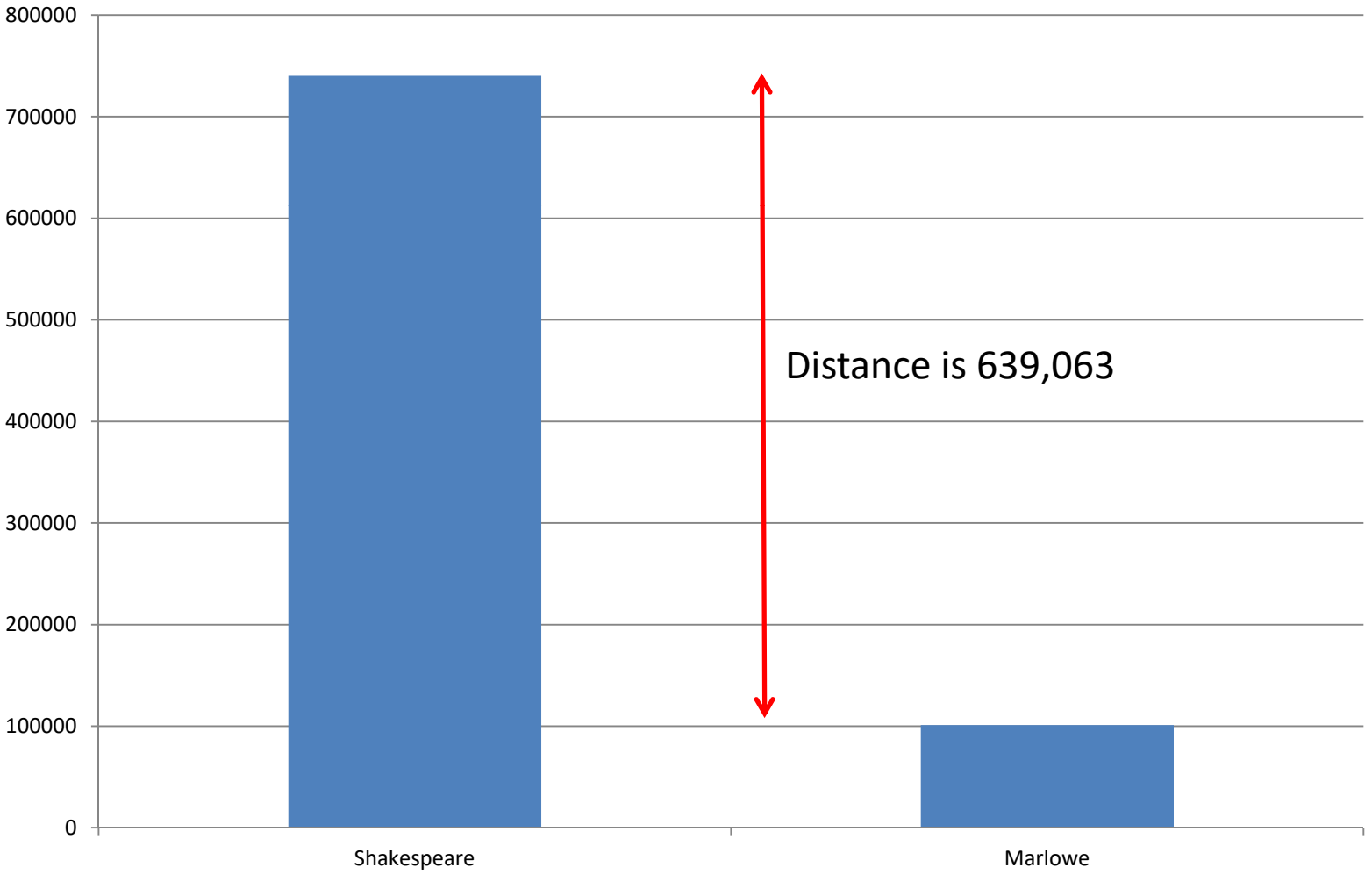
Distance is the absolute of subtraction:

| 4 - 7 | = 3 and also | 7 - 4 | = 3

"What is the distance between the number of words in Shakespeare's dramatic canon and the number of words in Christopher Marlowe's dramatic canon?"

Shakespeare = 740,209

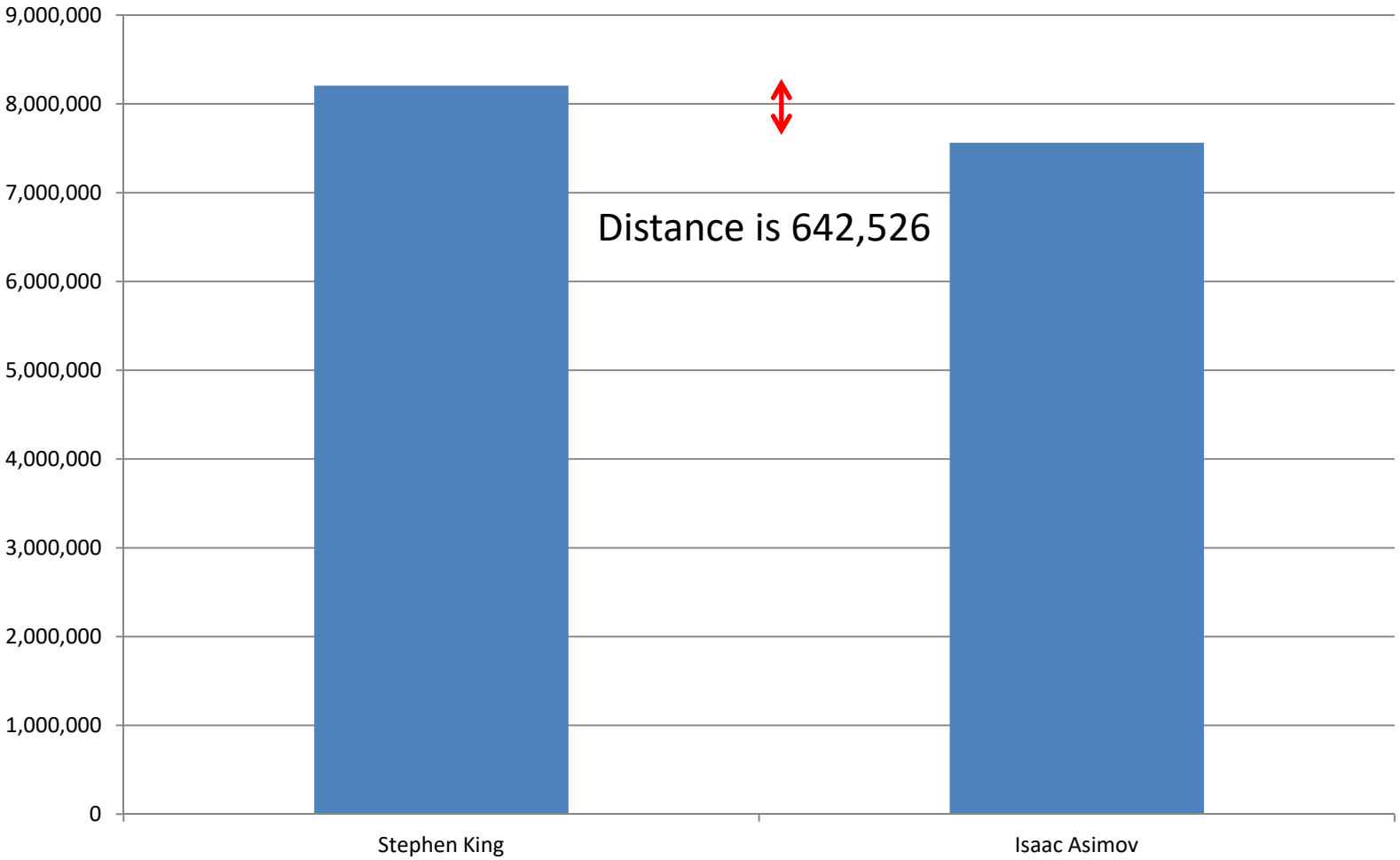
Marlowe = 101,146

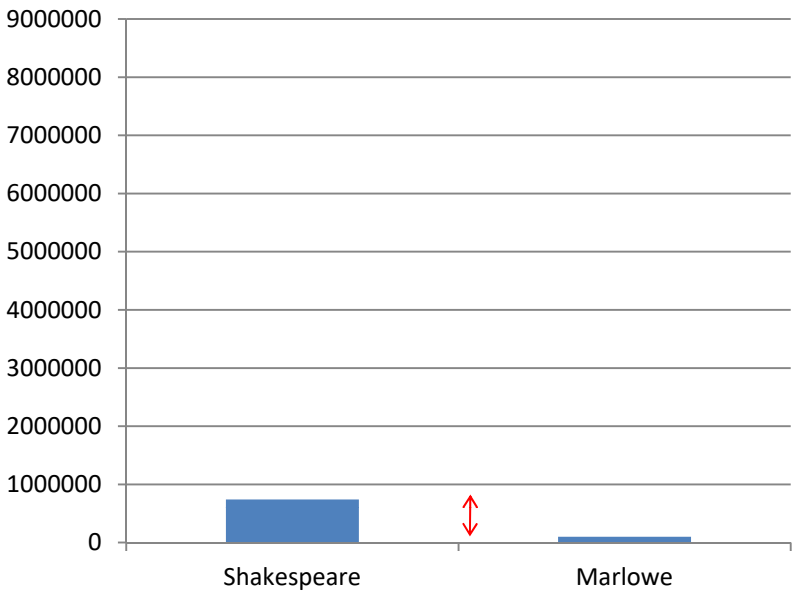


"What is the distance between the number of words in Stephen King's canon and the number of words in Isaac Asimov's canon?"

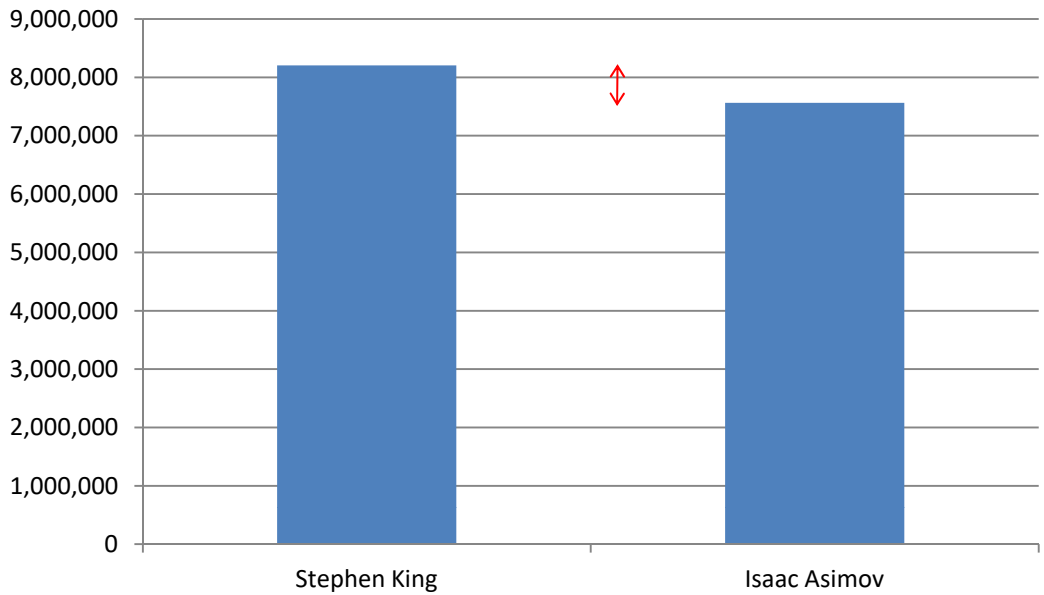
King = 8,205,051

Asimov = 7,562,525





Distance 639,063



Distance 642,526

Shakespeare : Marlowe

King : Asimov

?

÷ (division)

NOT

- (subtraction)

÷

NOT

–

740,209 ÷ 101,146 = 7.32

8,205,051 ÷ 7,562,525 = 1.08

George Chapman 12
All Fools
The Blind Beggar of Alexandria
Bussy D'Ambois
The Tragedy of Charles Duke of Byron
Caesar and Pompey
Sir Giles Goosecap
The Gentleman Usher
A Humorous Day's Mirth
May-Day
Monsieur D'Olive
The Revenge of Bussy D'Ambois
The Widow's Tears

John Fletcher 15
Bonduca
Monsieur Thomas
Rule a Wife and Have a Wife
The Chances
The Faithful Shepherdess
The Humourous Lieutenant
The Island Princess
The Loyal Subject
The Mad Lover
The Pilgrim
The Wild Goose Chase.txt
The Woman's Prize
Valentinian
Women Pleased
A Wife for a Month

Robert Greene 4
Alphonsus, King of Aragon
Friar Bacon and Friar Bungay
Orlando Furioso
James IV

Ben Jonson 16
The Alchemist
Bartholomew Fair
Catiline's Conspiracy
Cynthia's Revels
The Devil is an Ass
Every Man In his Humour
Every Man Out of his Humour
Epicoene
The Magnetic Lady
The New Inn
Poetaster
Sejanus's Fall
The Sad Shepherd
The Staple of News
The Tale of a Tub
Volpone

Christopher Marlowe 6
1 Tamburlaine
2 Tamburlaine
Edward II
Doctor Faustus
The Jew of Malta
The Massacre at Paris

Thomas Middleton 16
The Second Maiden's Tragedy
A Chaste Maid in Cheapside
A Game at Chess
Hengist King of Kent
More Dissemblers Besides Women
Michaelmas Term
A Mad World My Masters
No Wit No Help Like a Woman's
The Phoenix
The Puritan
The Revenger's Tragedy
A Trick to Catch the Old One
Women Beware Women
The Widow
The Witch
Your Five Gallants

Peele 5
The Battle of Alcazar
The Love of King David and Fair Bethsabe
Edward I
The Old Wives Tale
The Arraignment of Paris

Shakespeare 27
1 Henry IV
2 Henry IV
Much Ado About Nothing
Antony and Cleopatra
All's Well that Ends Well
As You Like It
Coriolanus
Cymbeline
The Comedy of Errors
Henry V
Hamlet
Julius Caesar
Love's Labour's Lost
King Lear
A Midsummer Night's Dream
The Merchant of Venice
Othello
Richard II
Richard III
Romeo and Juliet
The Taming of the Shrew
The Two Gentlemen of Verona
The Tempest
Twelfth Night
Troilus and Cressida
The Merry Wives of Windsor
The Winter's Tale

$$\text{variety of language} = \frac{\text{number of word types}}{\text{number of word tokens}}$$

	types	tokens	types/tokens	tokens/types
George Chapman	17722	237604	0.075	13.41
John Fletcher	16700	339744	0.049	20.34
Robert Greene	8187	66967	0.122	8.18
Ben Jonson	26680	441301	0.06	16.54
Christopher Marlowe	10663	101506	0.105	9.52
Thomas Middleton	22025	332972	0.066	15.12
George Peele	9262	70662	0.131	7.63
Shakespeare	30216	638302	0.047	21.12

$$\text{variety of language} = \frac{\text{number of word types}}{\text{number of word tokens}}$$

	types	tokens	types/tokens	tokens/types
George Chapman	17722	237604	0.075	13.41
John Fletcher	16700	339744	0.049	20.34
Robert Greene	8187	66967	0.122	8.18
Ben Jonson	26680	441301	0.06	16.54
Christopher Marlowe	10663	101506	0.105	9.52
Thomas Middleton	22025	332972	0.066	15.12
George Peele	9262	70662	0.131	7.63
Shakespeare	30216	638302	0.047	21.12

$$\text{variety of language} = \frac{\text{number of word types}}{\text{number of word tokens}}$$

	types	tokens	types/tokens	tokens/types
George Chapman	17722	237604	0.075	13.41
John Fletcher	16700	339744	0.049	20.34
Robert Greene	8187	66967	0.122	8.18
Ben Jonson	26680	441301	0.06	16.54
Christopher Marlowe	10663	101506	0.105	9.52
Thomas Middleton	22025	332972	0.066	15.12
George Peele	9262	70662	0.131	7.63
Shakespeare	30216	638302	0.047	21.12

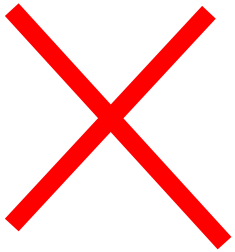
Chapman 12 **Fletcher 15** Greene 4 **Jonson 16**
Marlowe 6 Middleton 16 Peele 5 **Shakespeare 27**

	types	tokens	types/tokens	tokens/types
George Chapman	17722	237604	0.075	13.41
John Fletcher	16700	339744	0.049	20.34
Robert Greene	8187	66967	0.122	8.18
Ben Jonson	26680	441301	0.06	16.54
Christopher Marlowe	10663	101506	0.105	9.52
Thomas Middleton	22025	332972	0.066	15.12
George Peele	9262	70662	0.131	7.63
Shakespeare	30216	638302	0.047	21.12

Chapman 12
Fletcher 15
Greene 4
Jonson 16
Marlowe 6
Middleton 16
Peele 5
Shakespeare 27

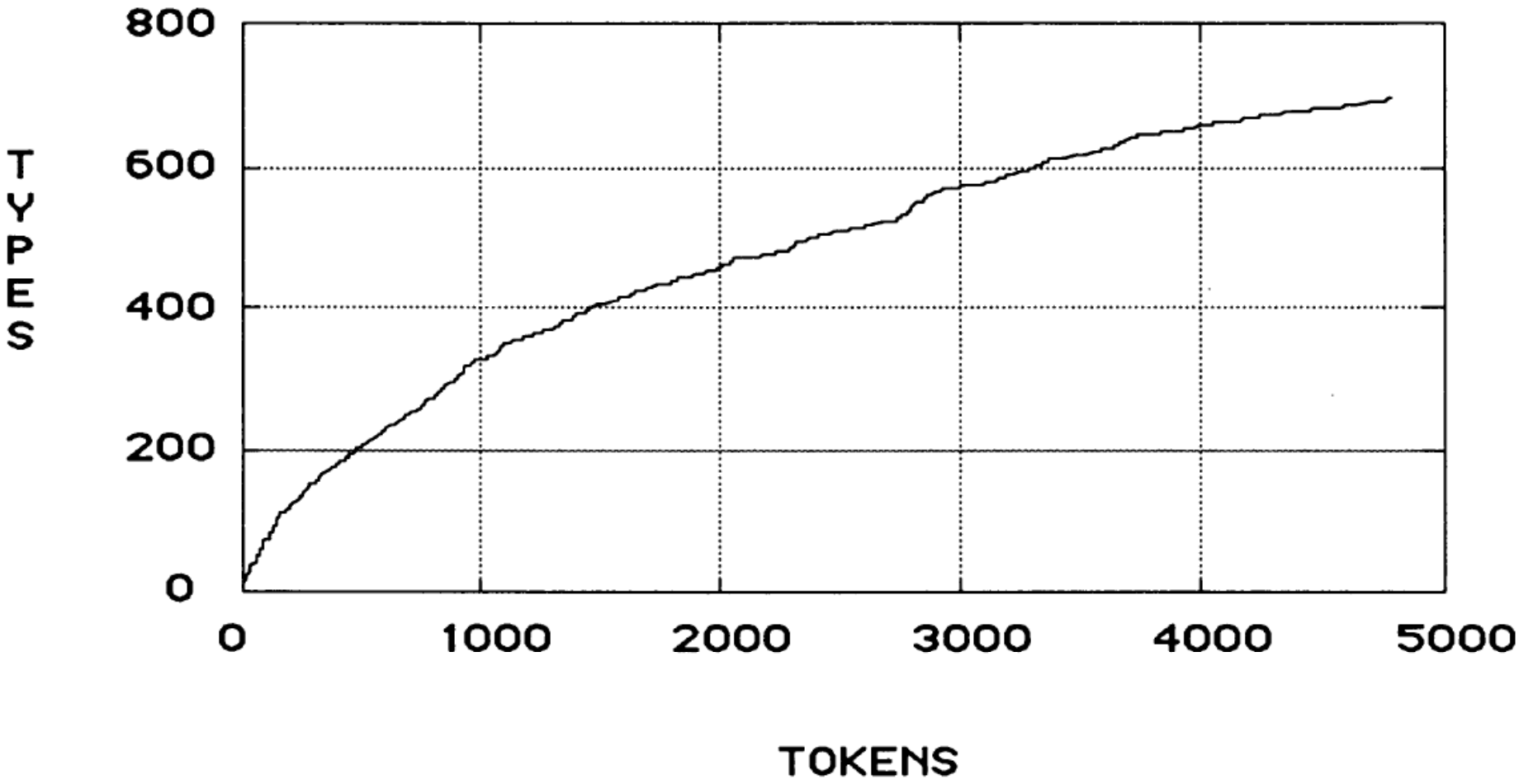
	types	tokens	types/tokens	tokens/types
George Chapman	17722	237604	0.075	13.41
John Fletcher	16700	339744	0.049	20.34
Robert Greene	8187	66967	0.122	8.18
Ben Jonson	26680	441301	0.06	16.54
Christopher Marlowe	10663	101506	0.105	9.52
Thomas Middleton	22025	332972	0.066	15.12
George Peele	9262	70662	0.131	7.63
Shakespeare	30216	638302	0.047	21.12

$$\text{variety of language} = \frac{\text{number of word types}}{\text{number of word tokens}}$$



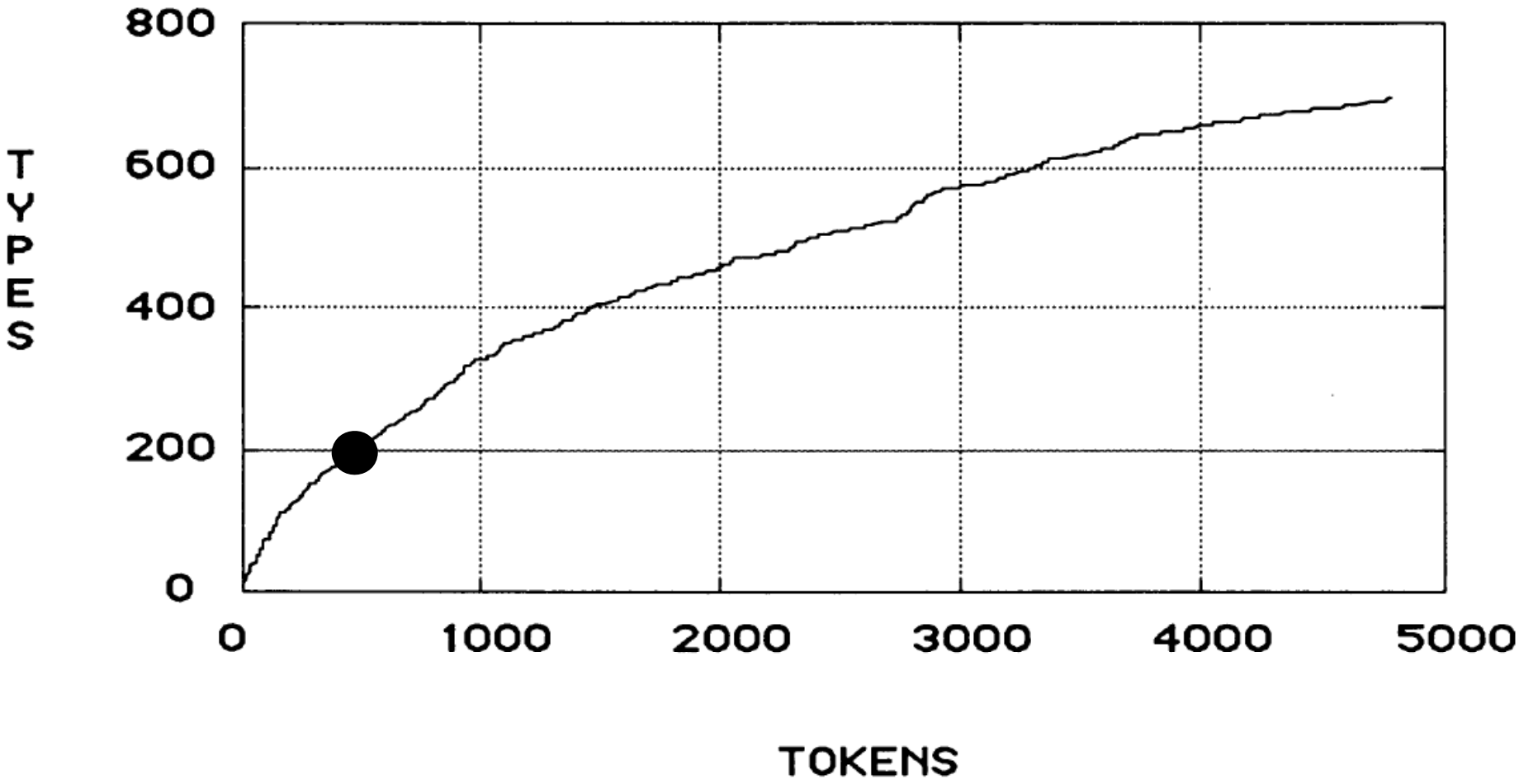
	types	tokens	types/tokens	tokens/types
George Chapman	17722	237604	0.075	13.41
John Fletcher	16700	339744	0.049	20.34
Robert Greene	8187	66967	0.122	8.18
Ben Jonson	26680	441301	0.06	16.54
Christopher Marlowe	10663	101506	0.105	9.52
Thomas Middleton	22025	332972	0.066	15.12
George Peele	9262	70662	0.131	7.63
Shakespeare	30216	638302	0.047	21.12

Fig. 2. Plot of Types Versus Tokens for "Macbeth" in Basic English



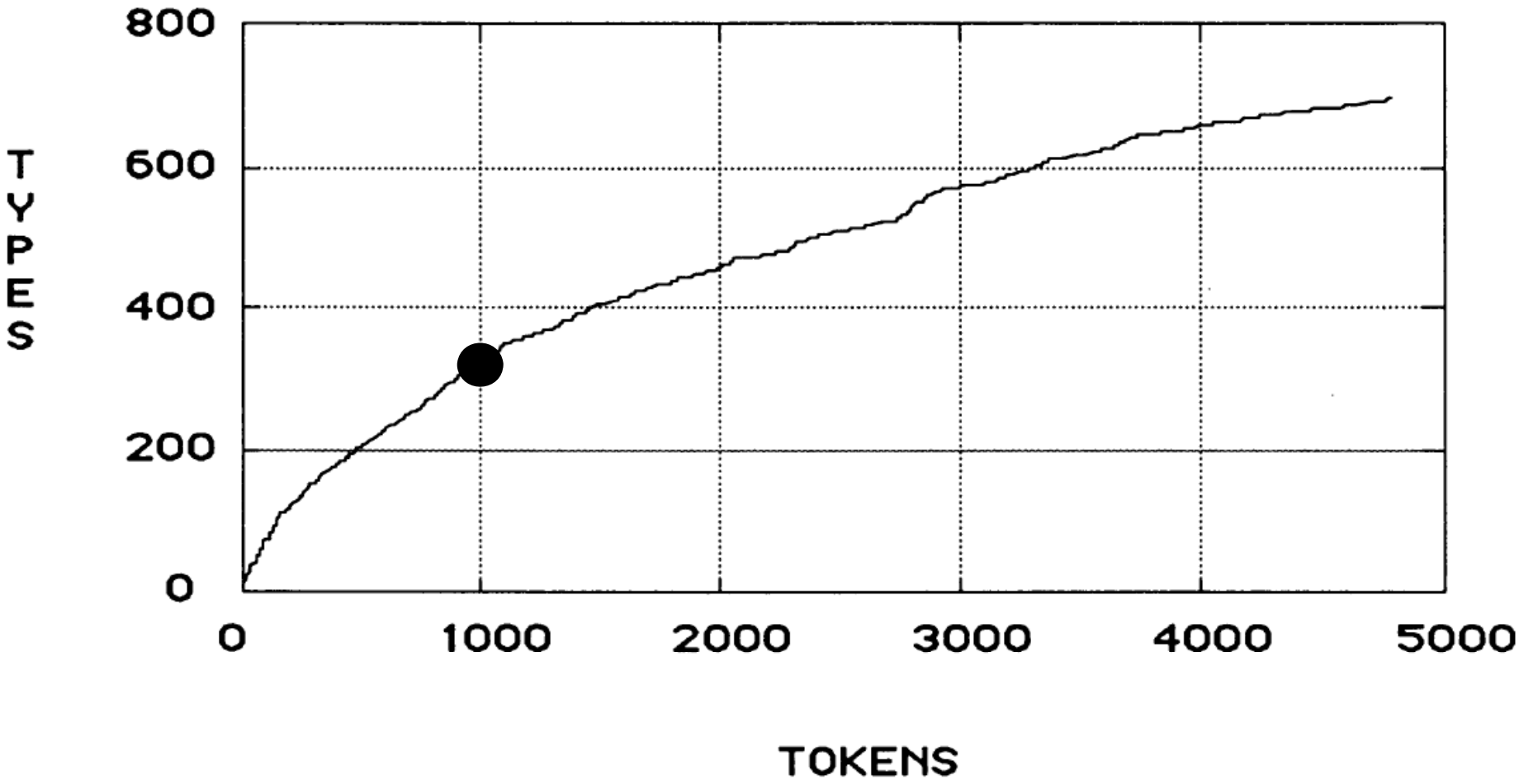
From Gilbert Youmans "Measuring lexical style and competence: The type-token vocabulary curve" *Style* 24 (1990): 584-599 (p. 588)

Fig. 2. Plot of Types Versus Tokens for "Macbeth" in Basic English



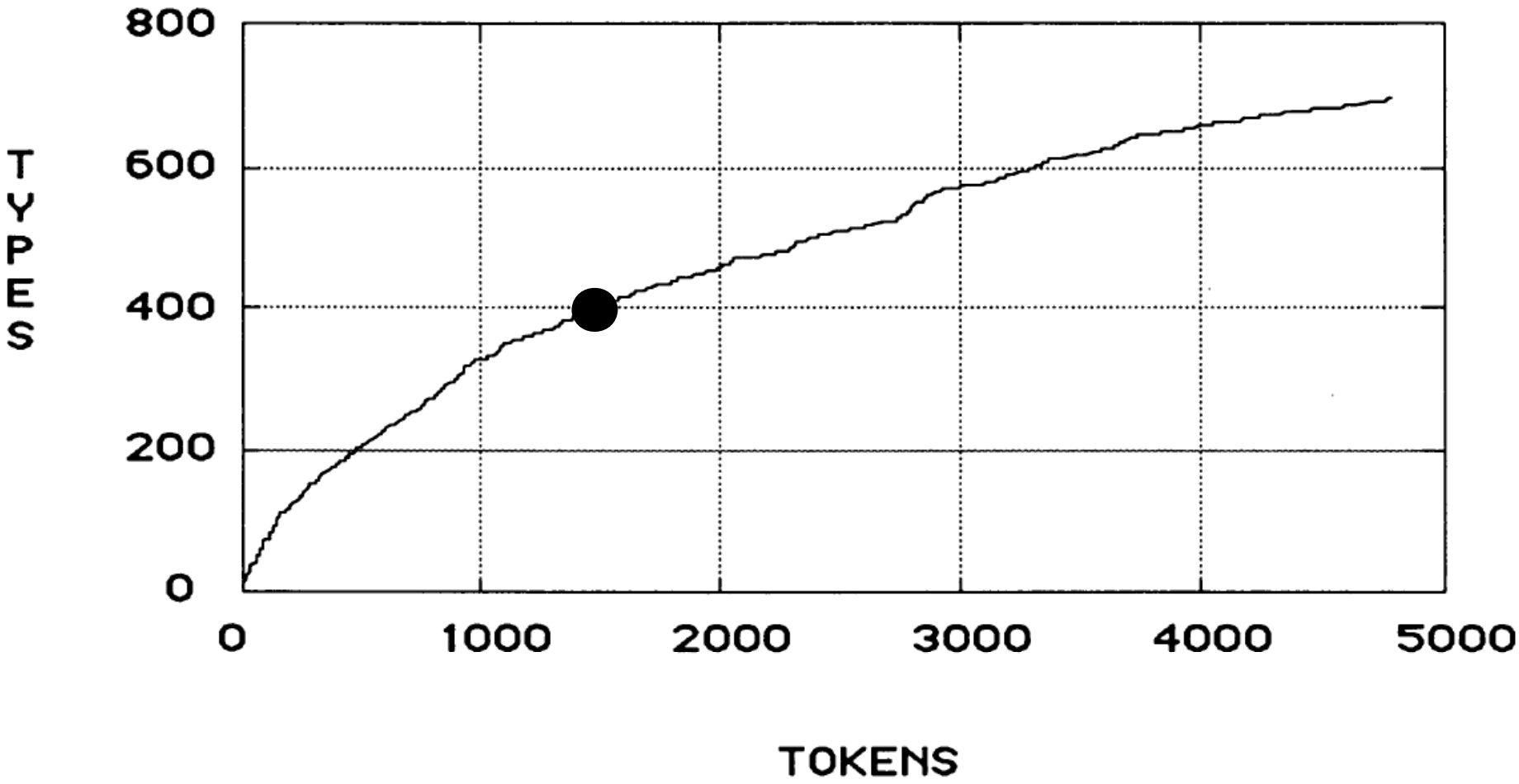
From Gilbert Youmans "Measuring lexical style and competence: The type-token vocabulary curve" *Style* 24 (1990): 584-599 (p. 588)

Fig. 2. Plot of Types Versus Tokens for "Macbeth" in Basic English



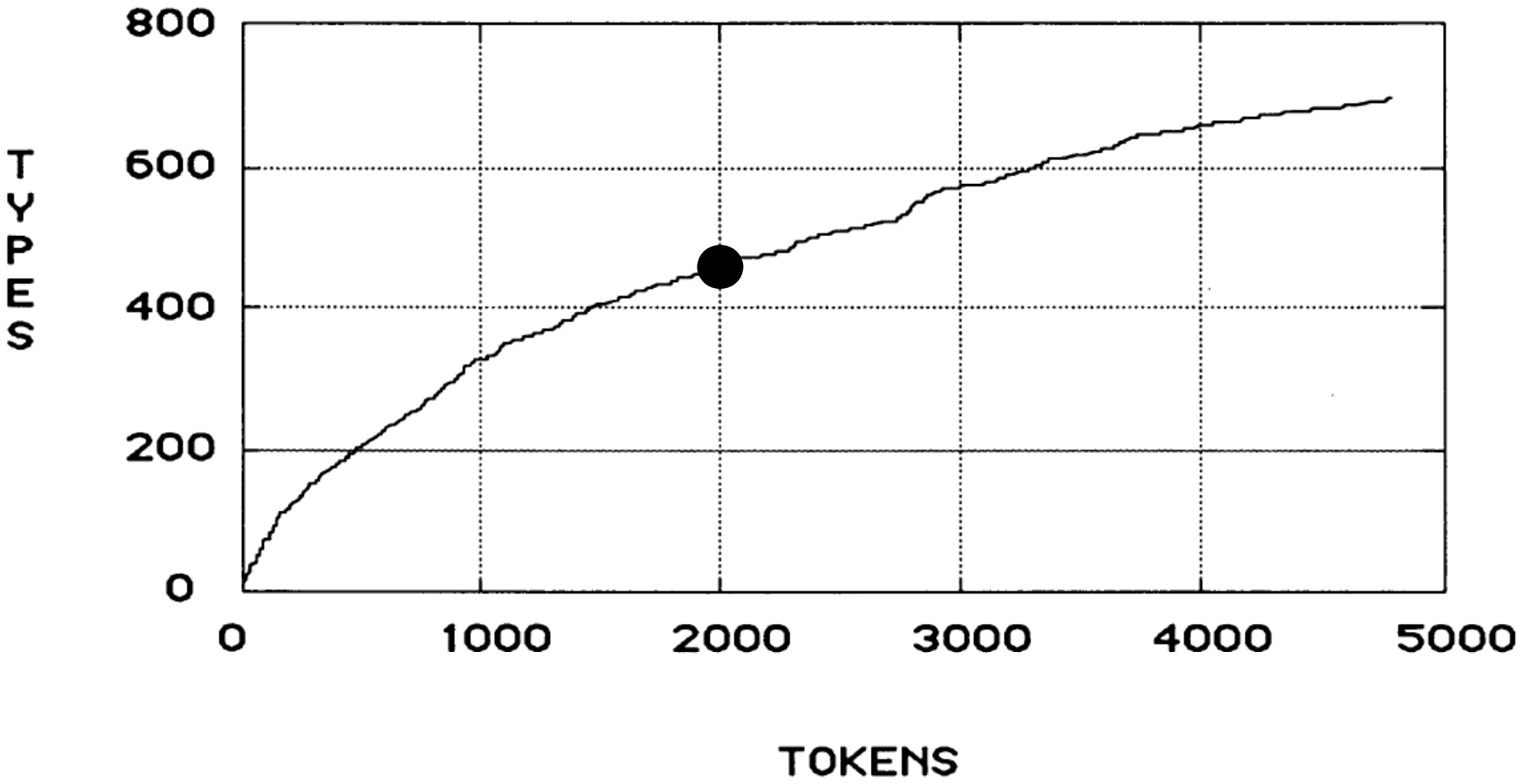
From Gilbert Youmans "Measuring lexical style and competence: The type-token vocabulary curve" *Style* 24 (1990): 584-599 (p. 588)

Fig. 2. Plot of Types Versus Tokens for "Macbeth" in Basic English



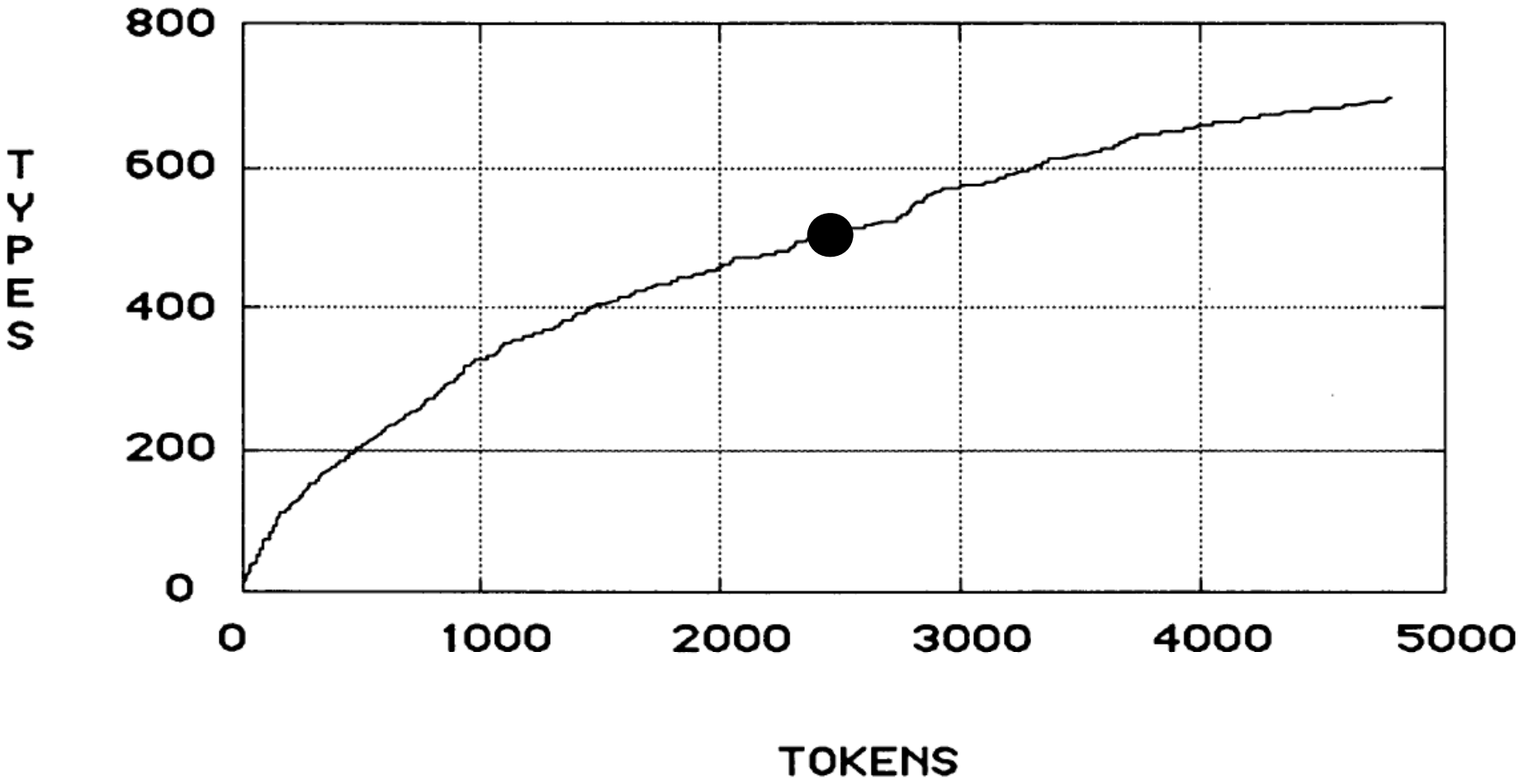
From Gilbert Youmans "Measuring lexical style and competence: The type-token vocabulary curve" *Style* 24 (1990): 584-599 (p. 588)

Fig. 2. Plot of Types Versus Tokens for "Macbeth" in Basic English



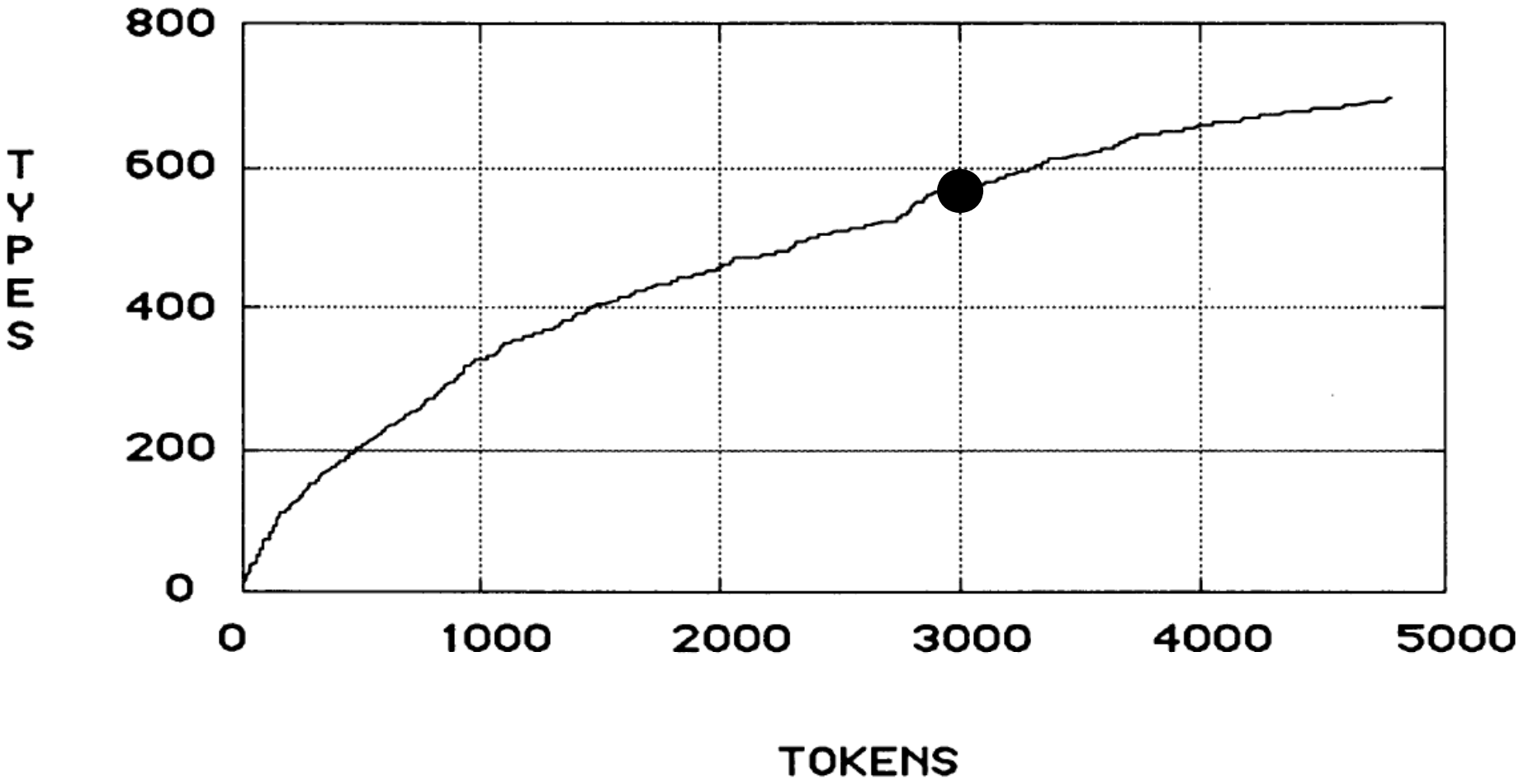
From Gilbert Youmans "Measuring lexical style and competence: The type-token vocabulary curve" *Style* 24 (1990): 584-599 (p. 588)

Fig. 2. Plot of Types Versus Tokens for "Macbeth" in Basic English



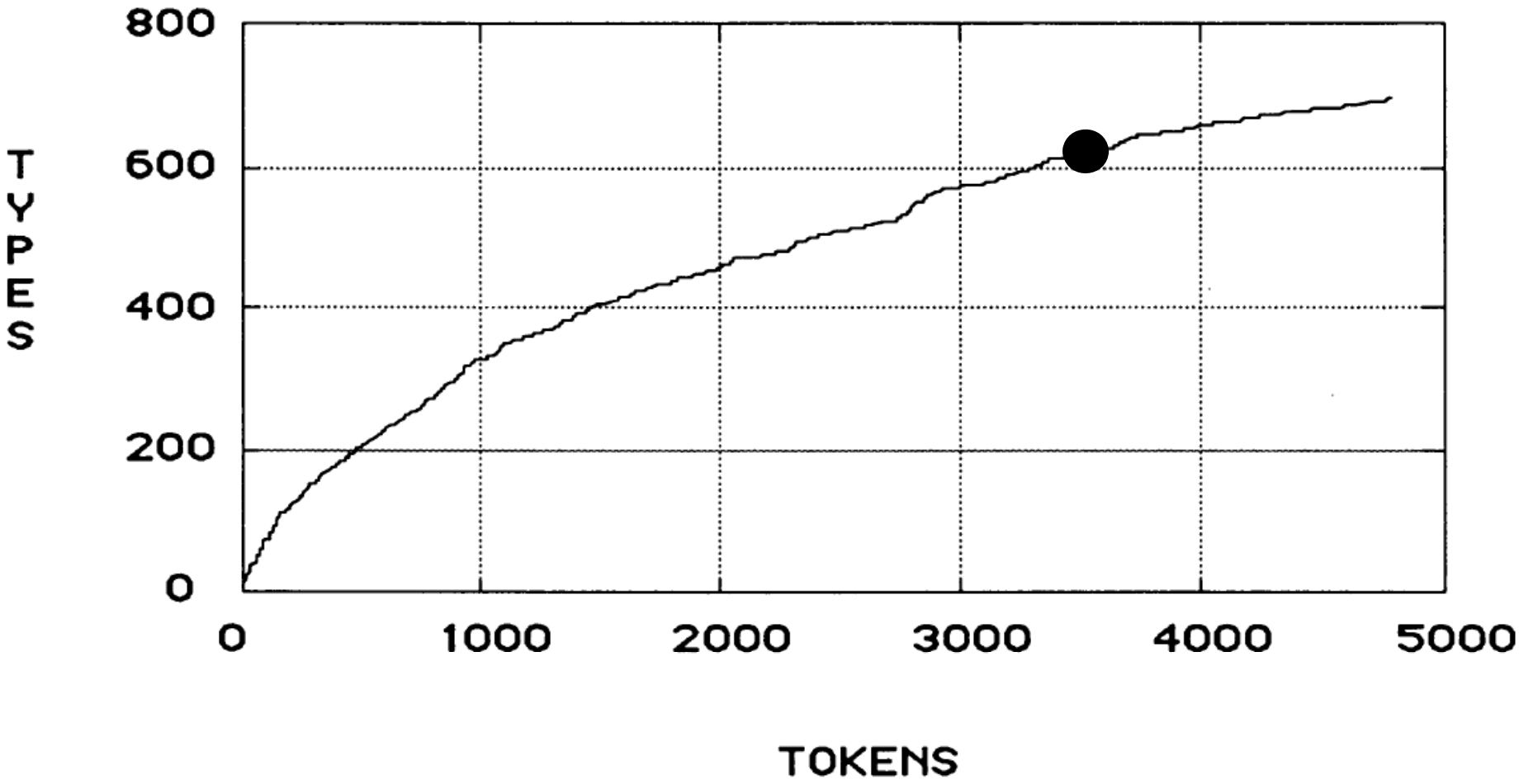
From Gilbert Youmans "Measuring lexical style and competence: The type-token vocabulary curve" *Style* 24 (1990): 584-599 (p. 588)

Fig. 2. Plot of Types Versus Tokens for "Macbeth" in Basic English



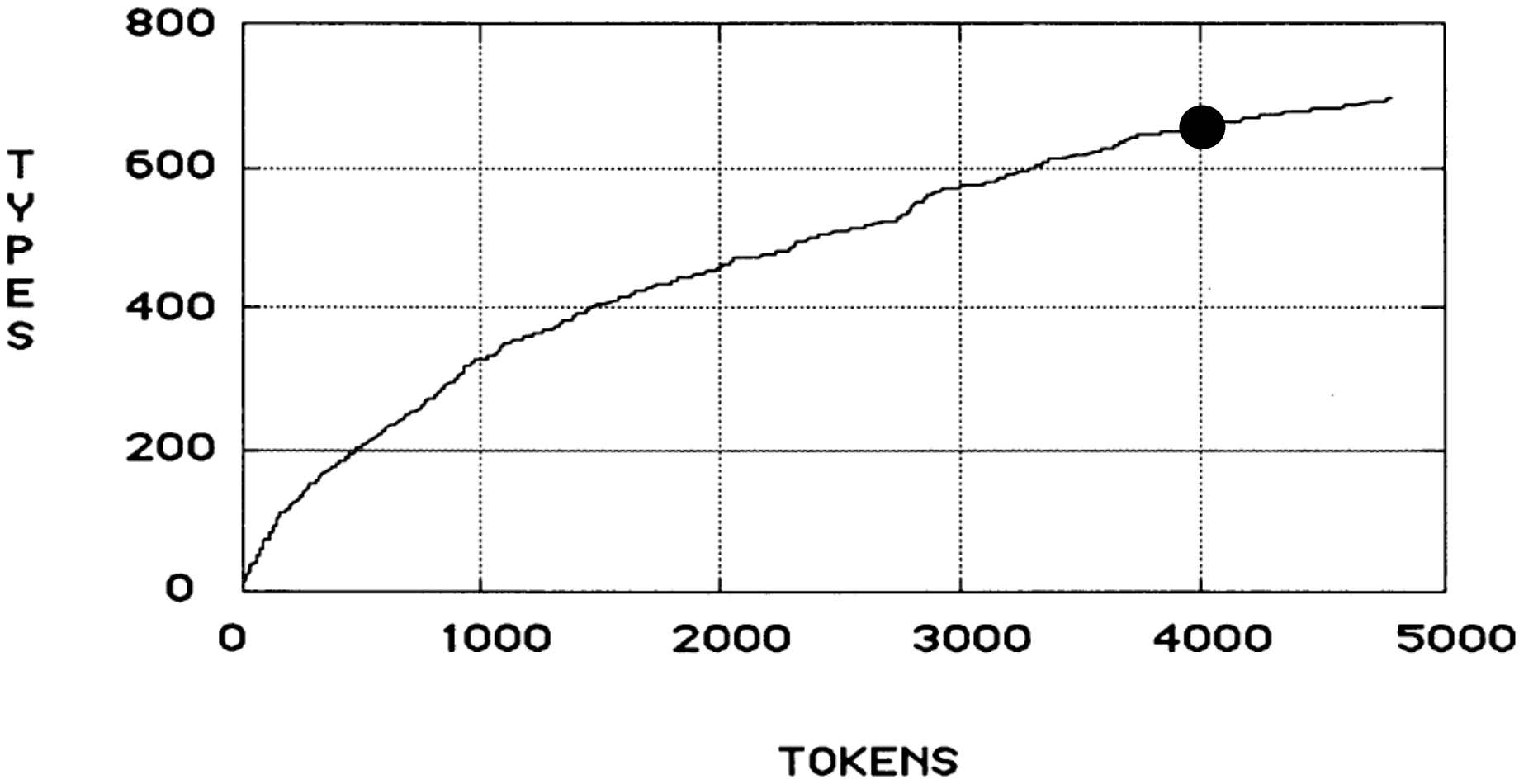
From Gilbert Youmans "Measuring lexical style and competence: The type-token vocabulary curve" *Style* 24 (1990): 584-599 (p. 588)

Fig. 2. Plot of Types Versus Tokens for "Macbeth" in Basic English



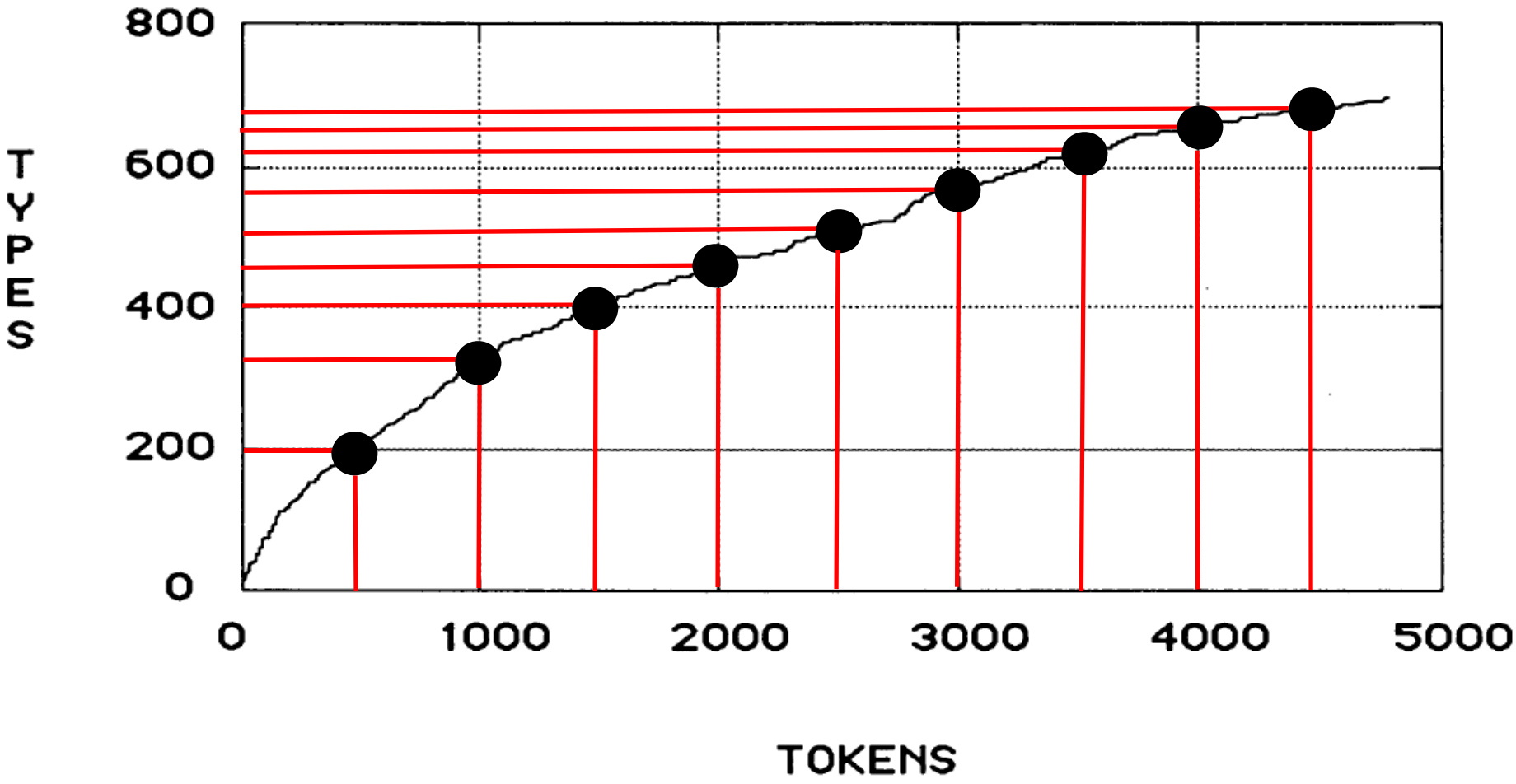
From Gilbert Youmans "Measuring lexical style and competence: The type-token vocabulary curve" *Style* 24 (1990): 584-599 (p. 588)

Fig. 2. Plot of Types Versus Tokens for "Macbeth" in Basic English



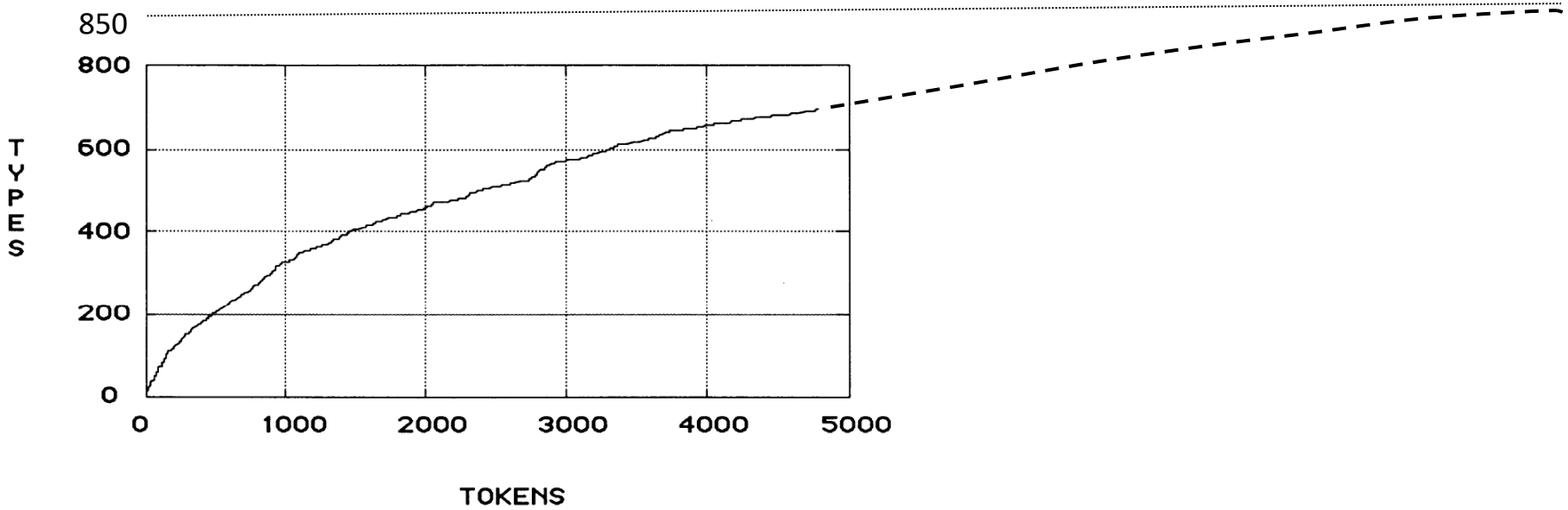
From Gilbert Youmans "Measuring lexical style and competence: The type-token vocabulary curve" *Style* 24 (1990): 584-599 (p. 588)

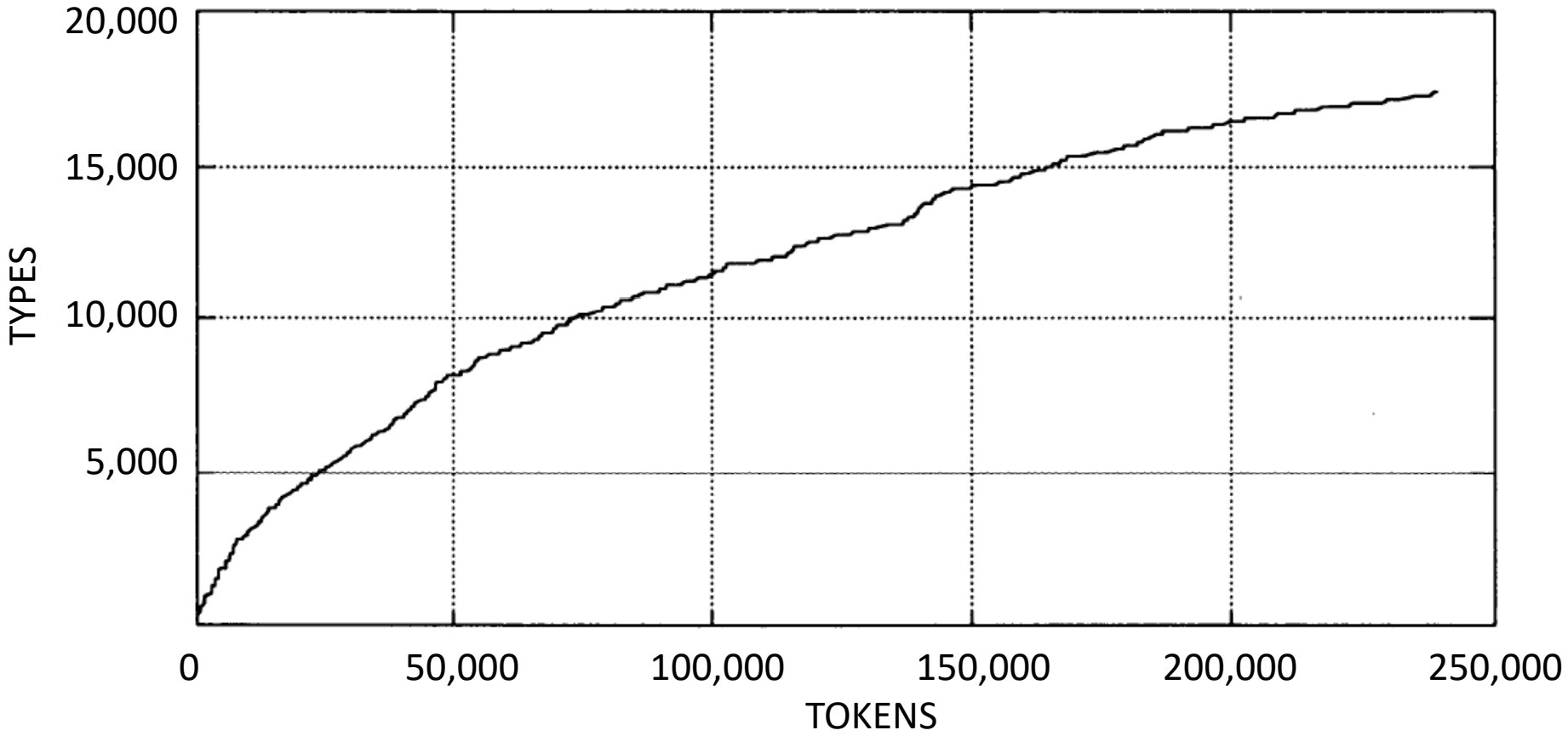
Fig. 2. Plot of Types Versus Tokens for "Macbeth" in Basic English

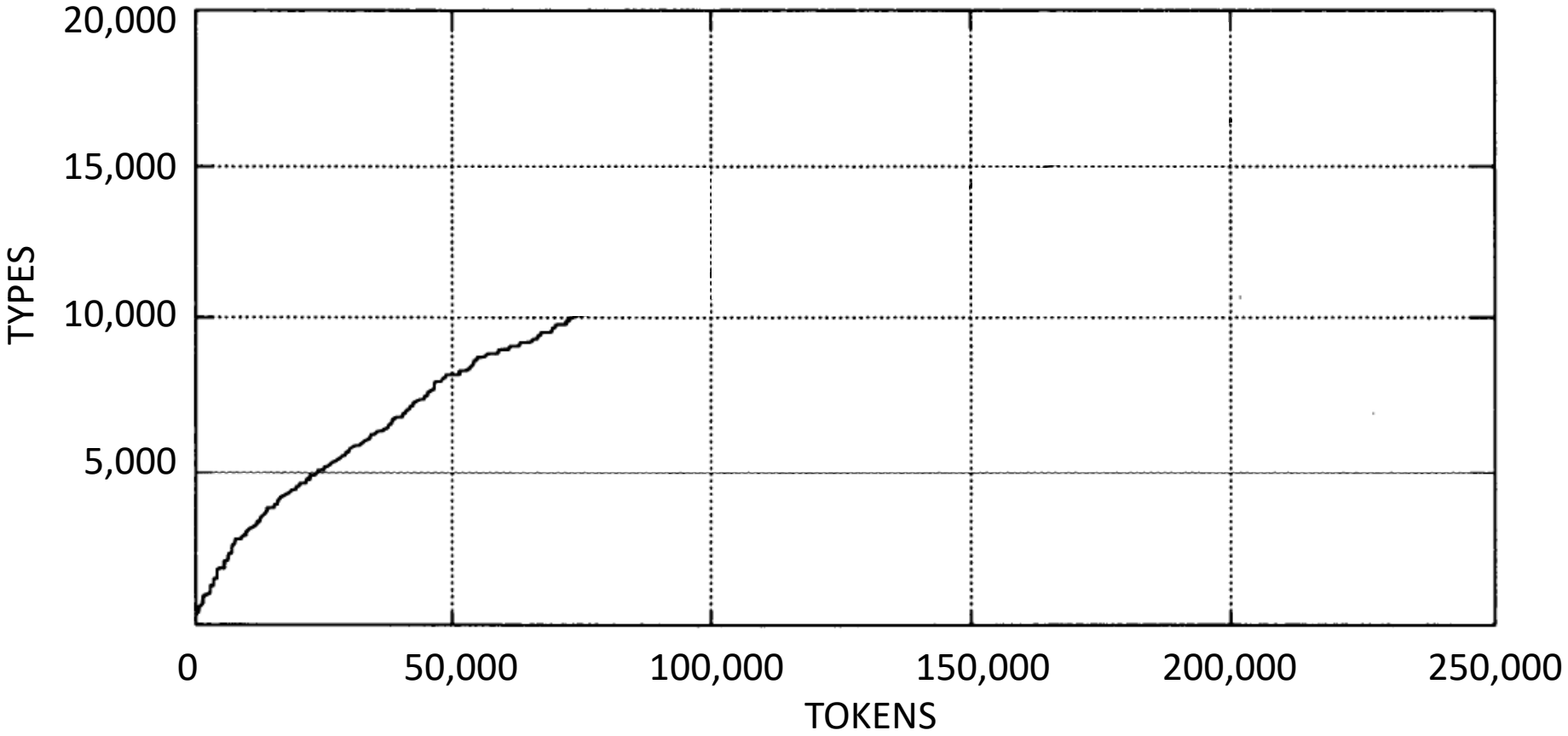


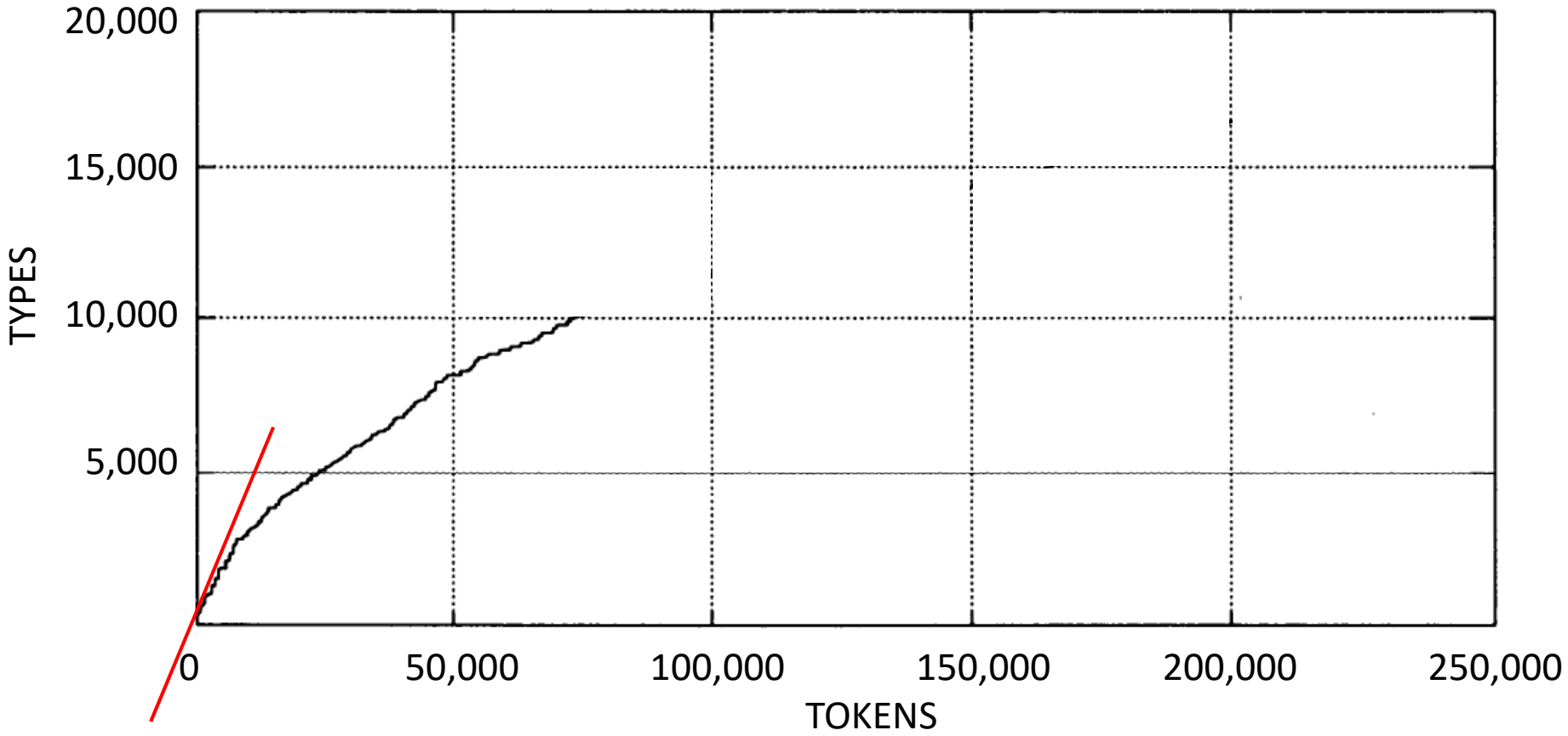
From Gilbert Youmans "Measuring lexical style and competence: The type-token vocabulary curve" *Style* 24 (1990): 584-599 (p. 588)

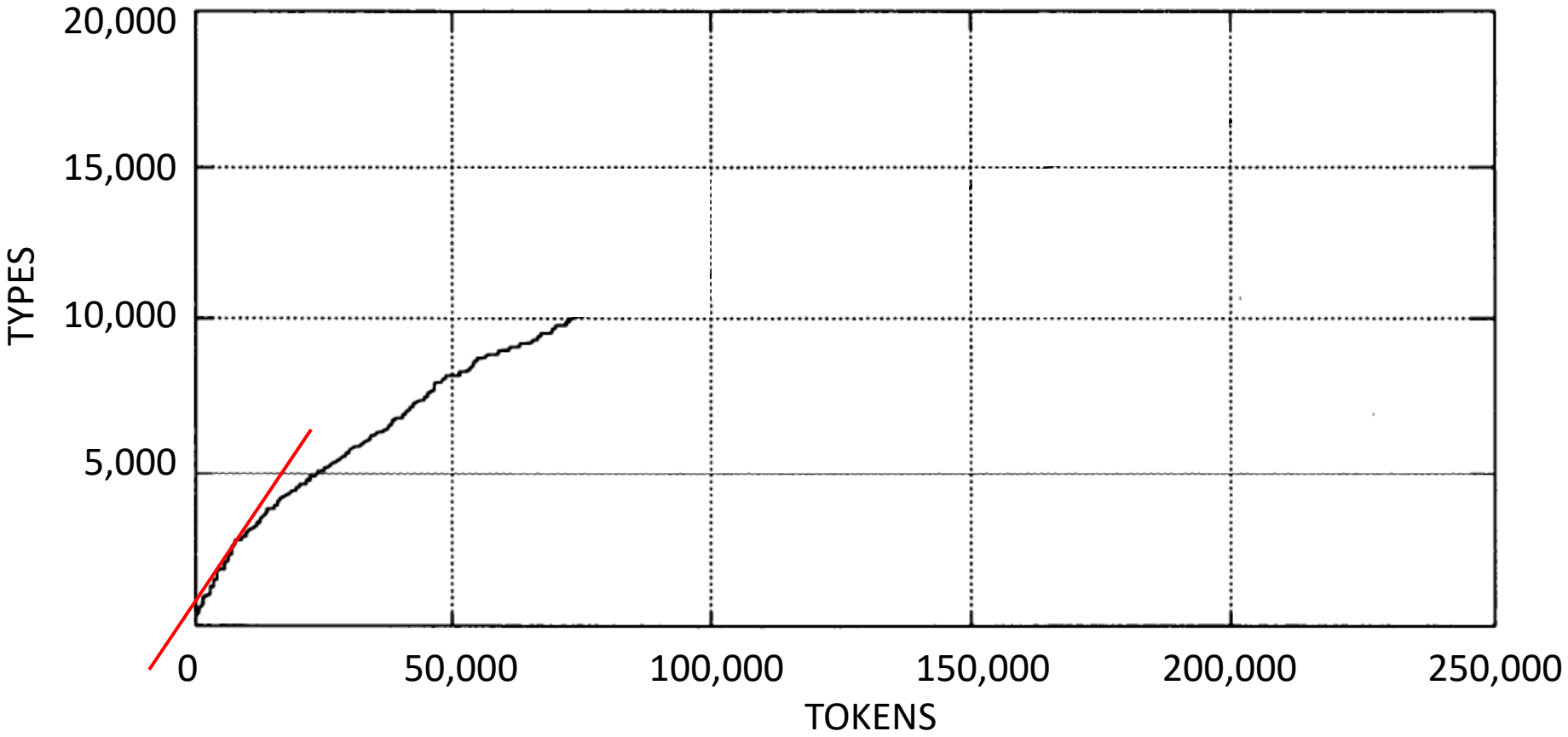
Fig. 2. Plot of Types Versus Tokens for "Macbeth" in Basic English

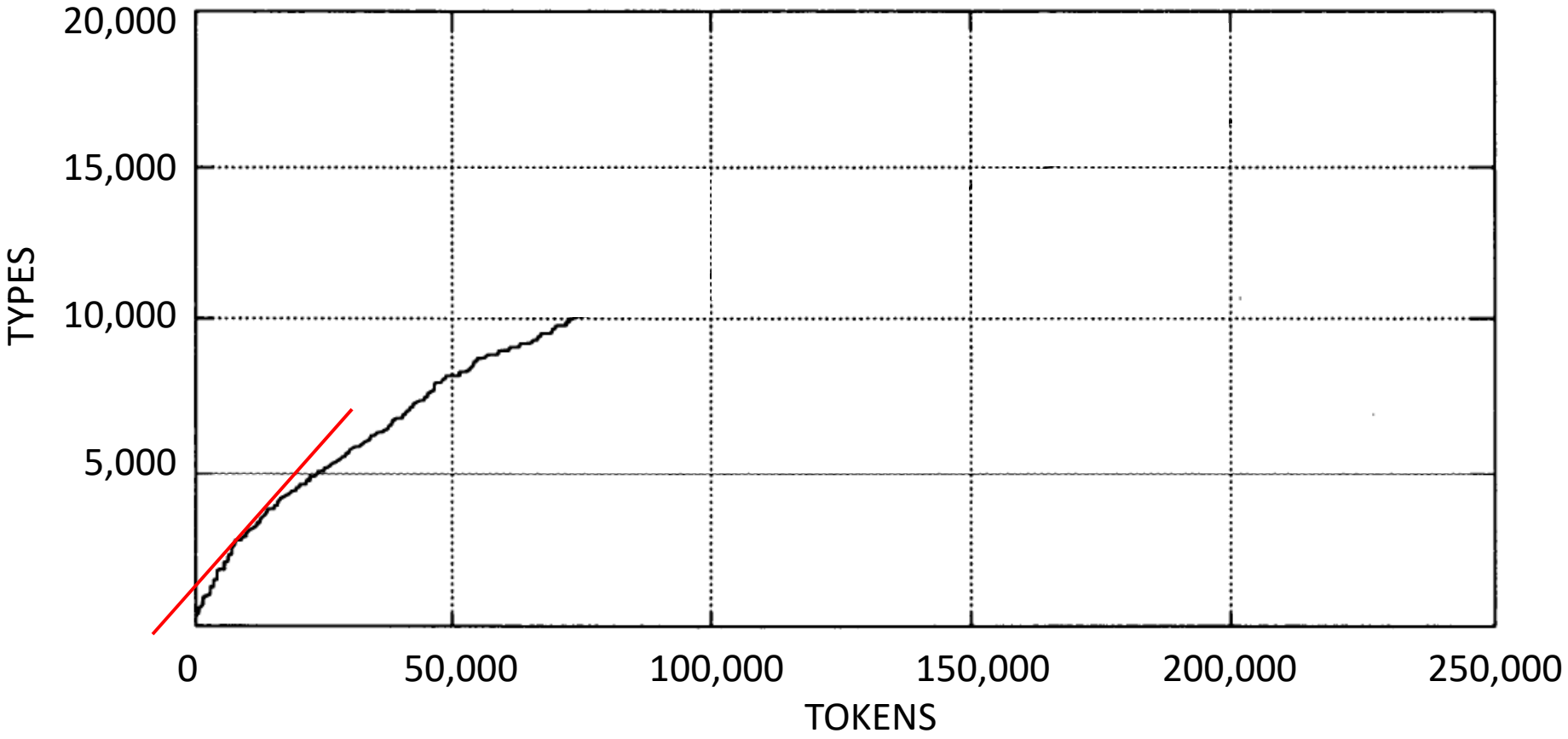


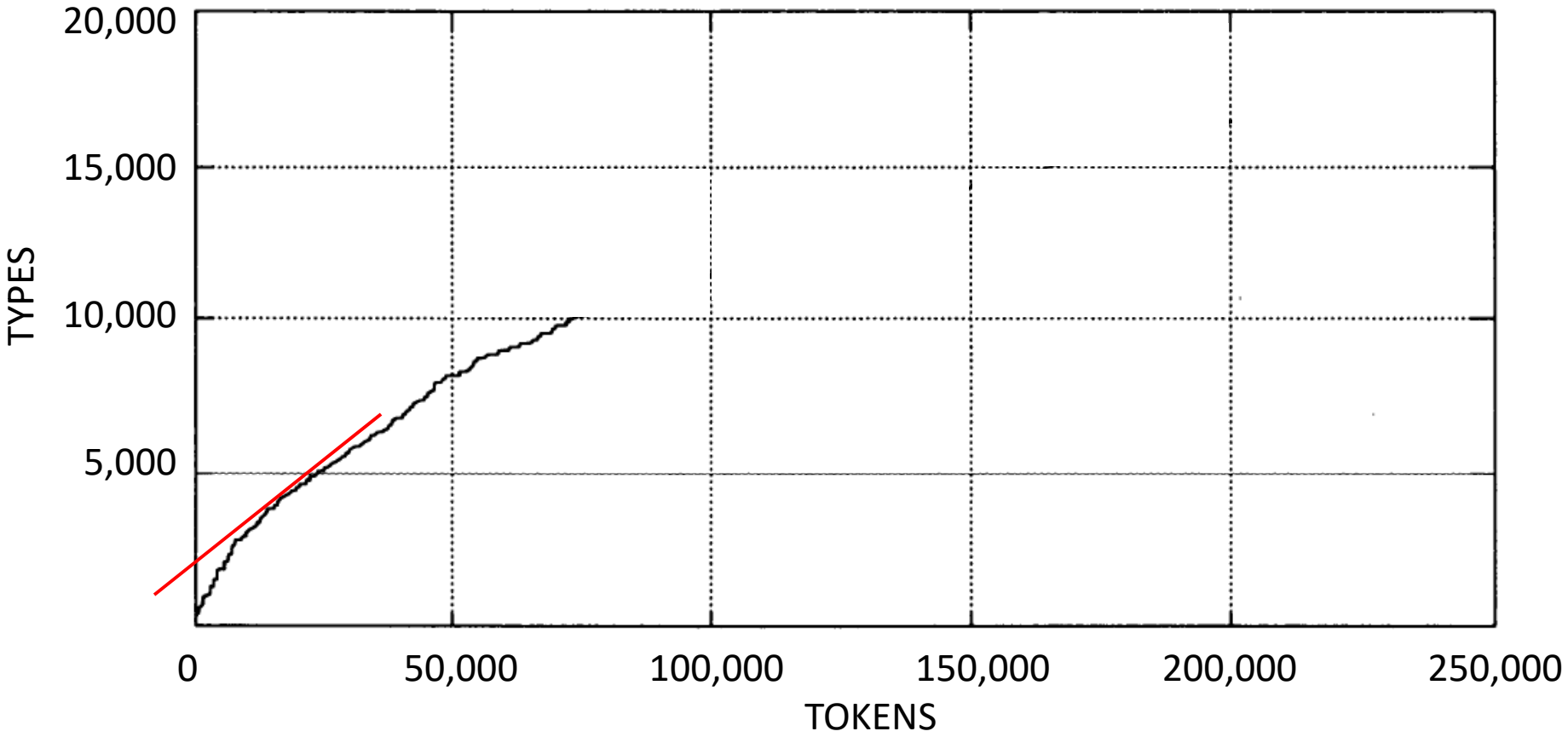

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

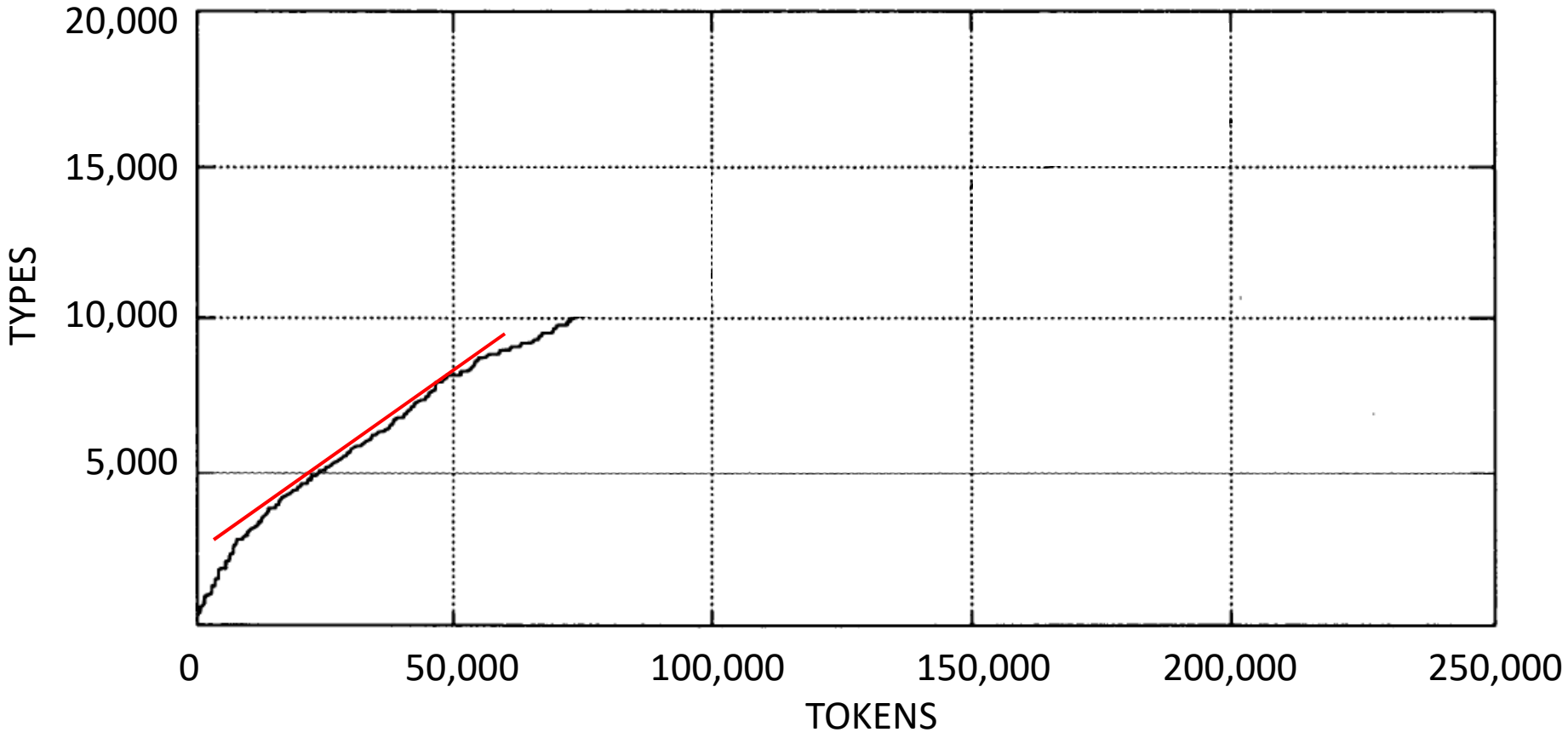

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

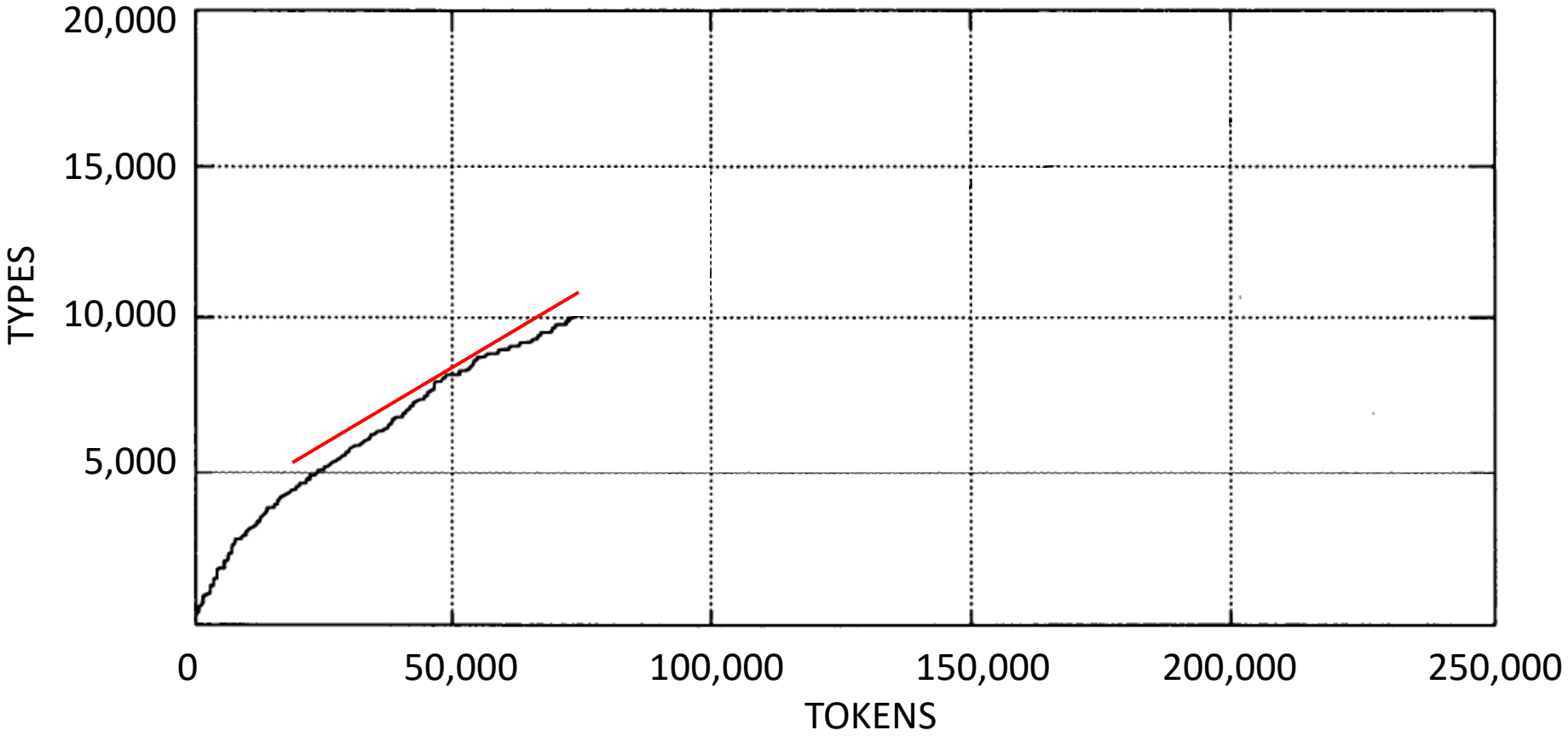

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

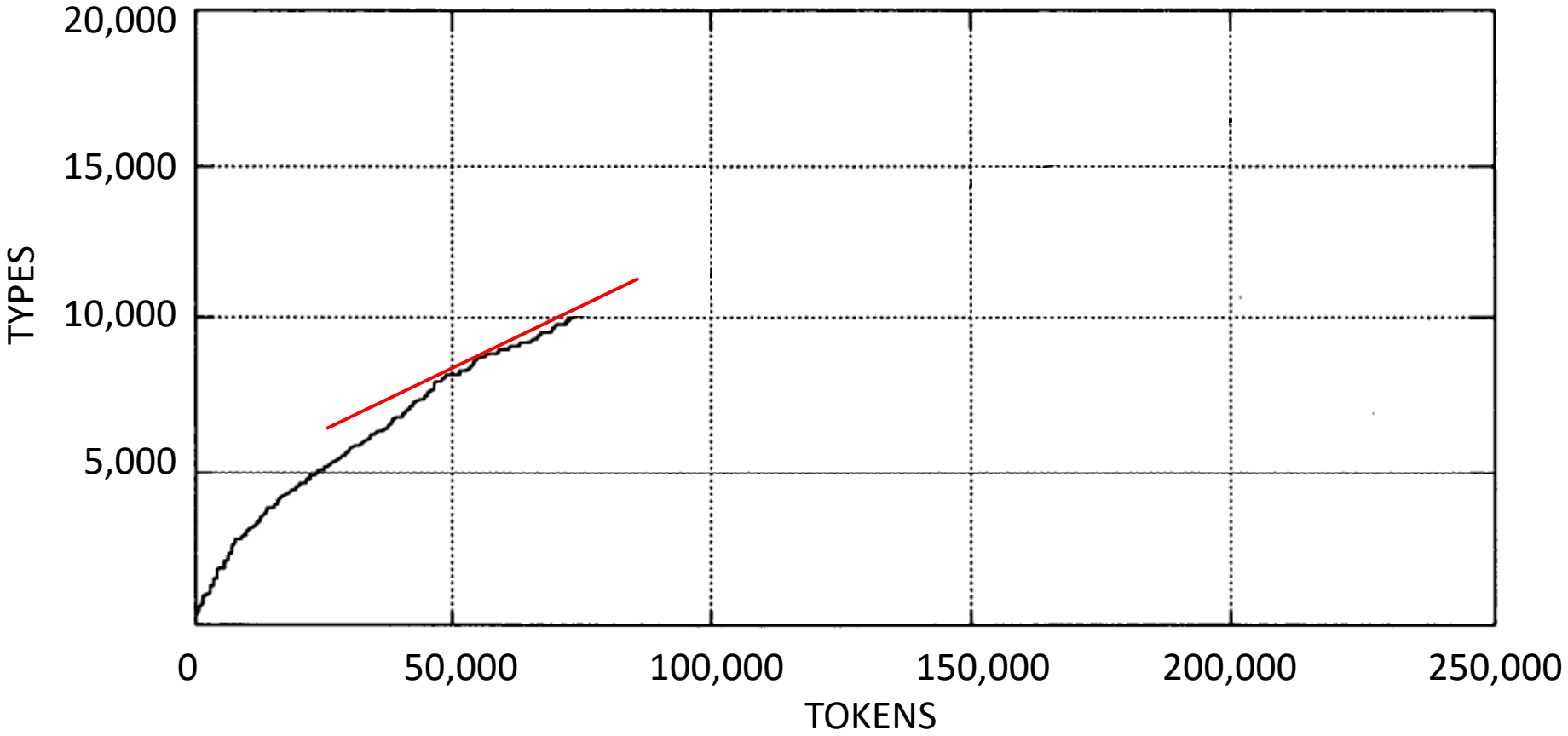

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

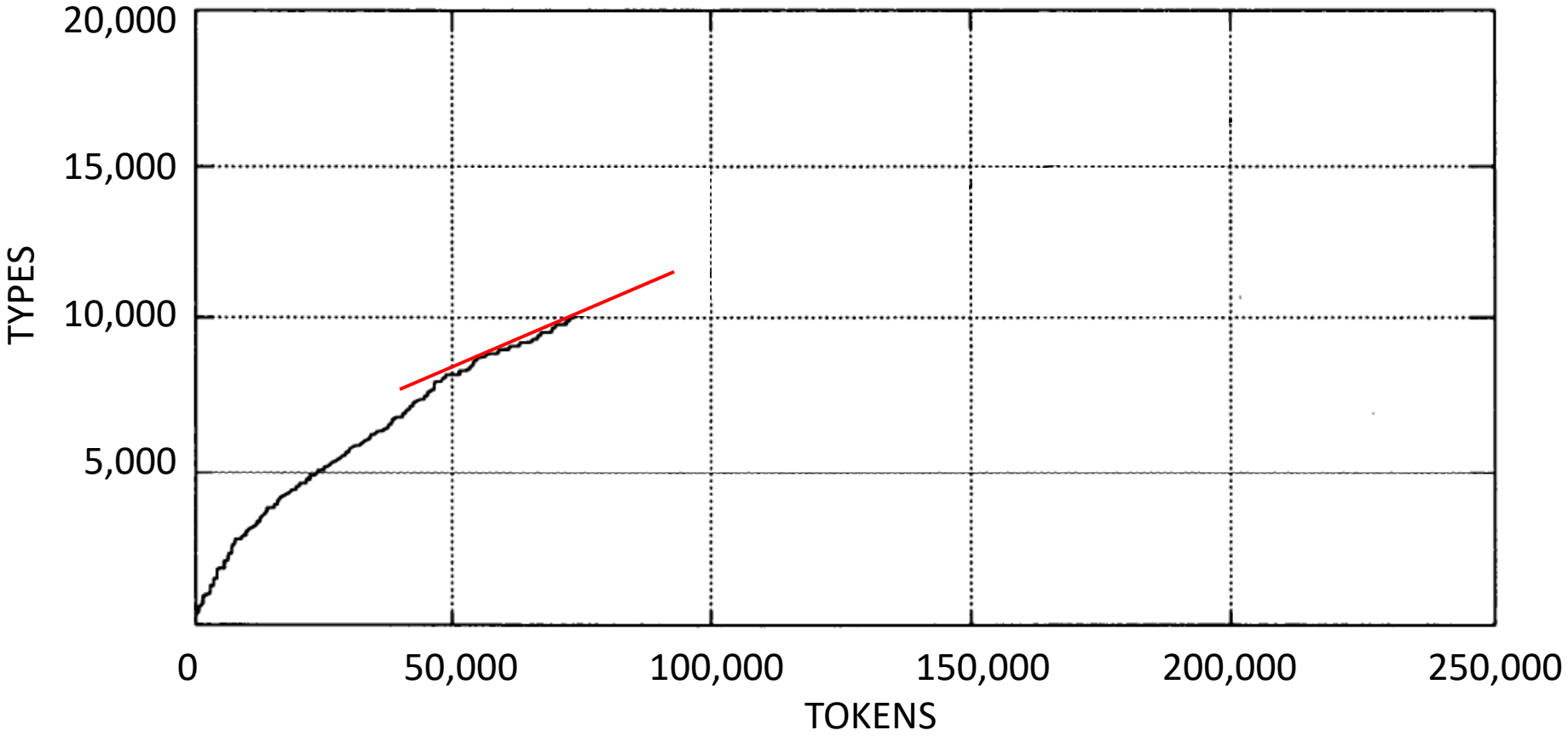

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

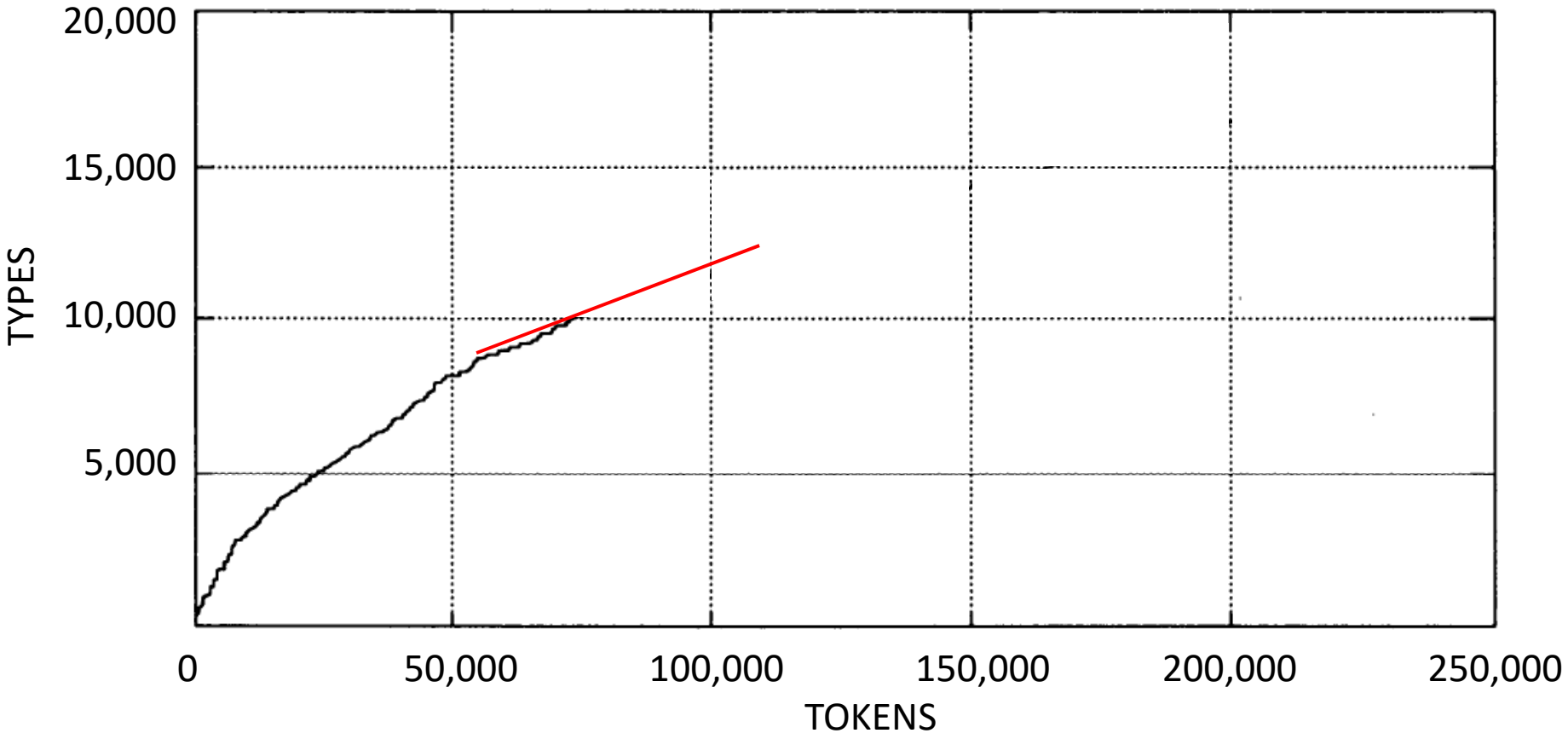

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

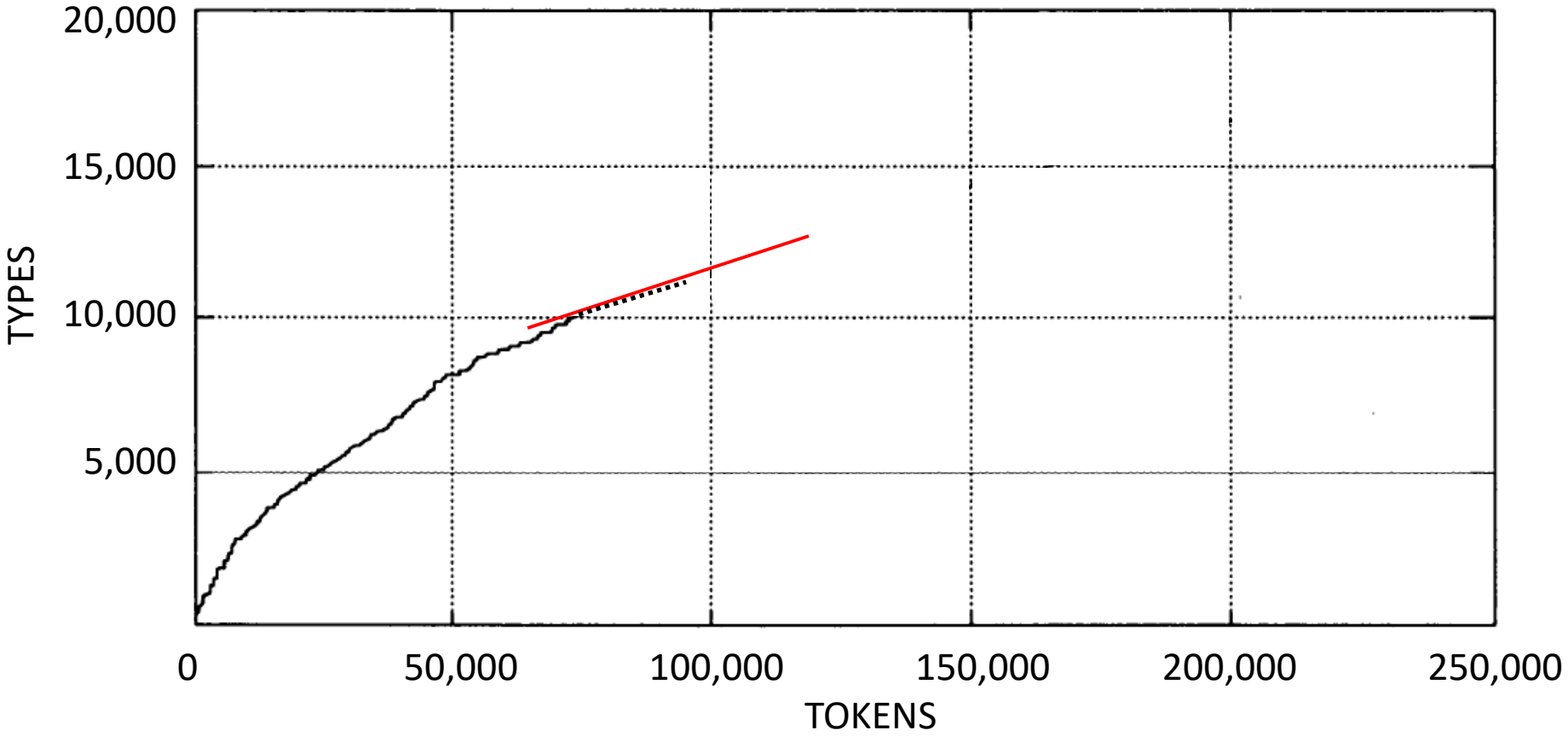

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

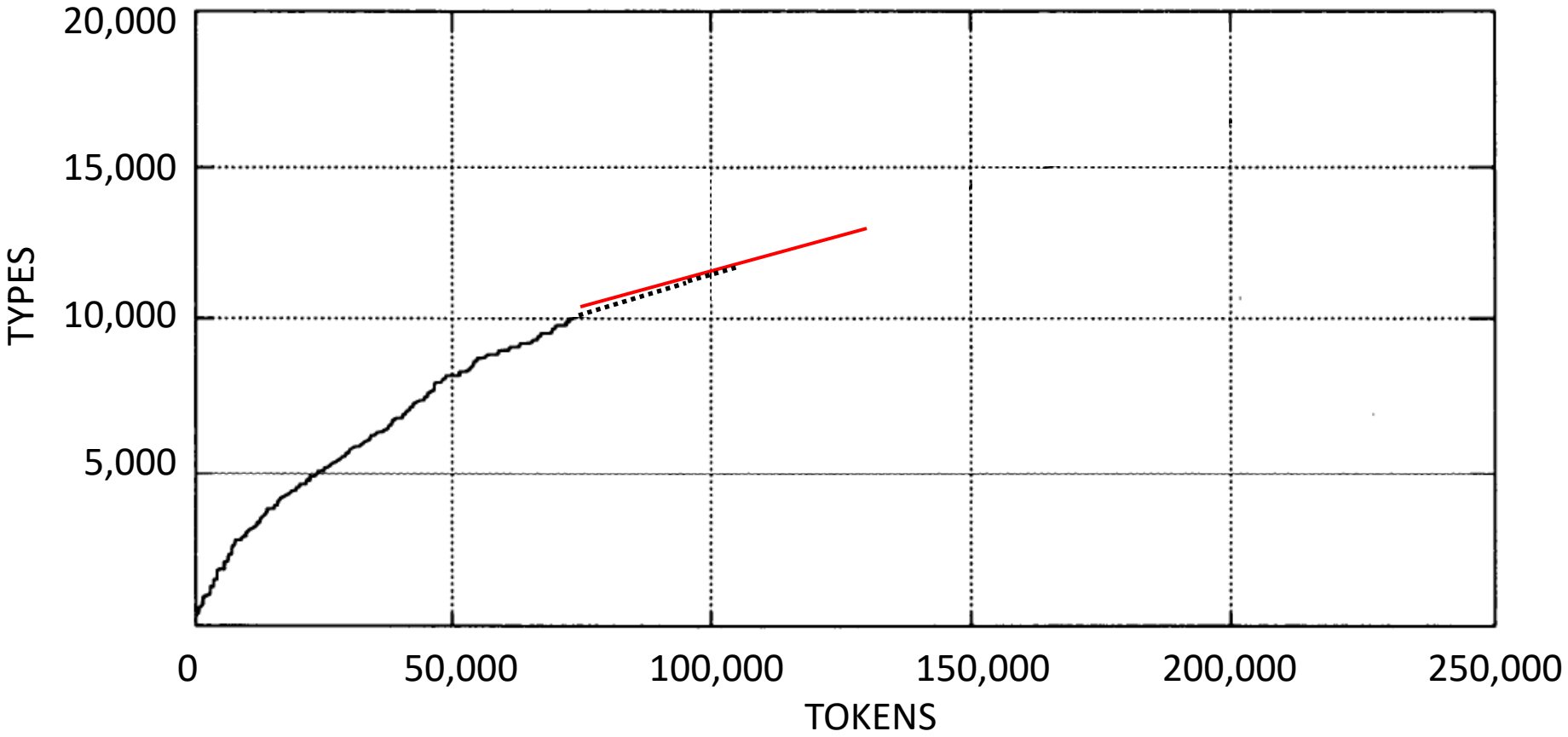

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

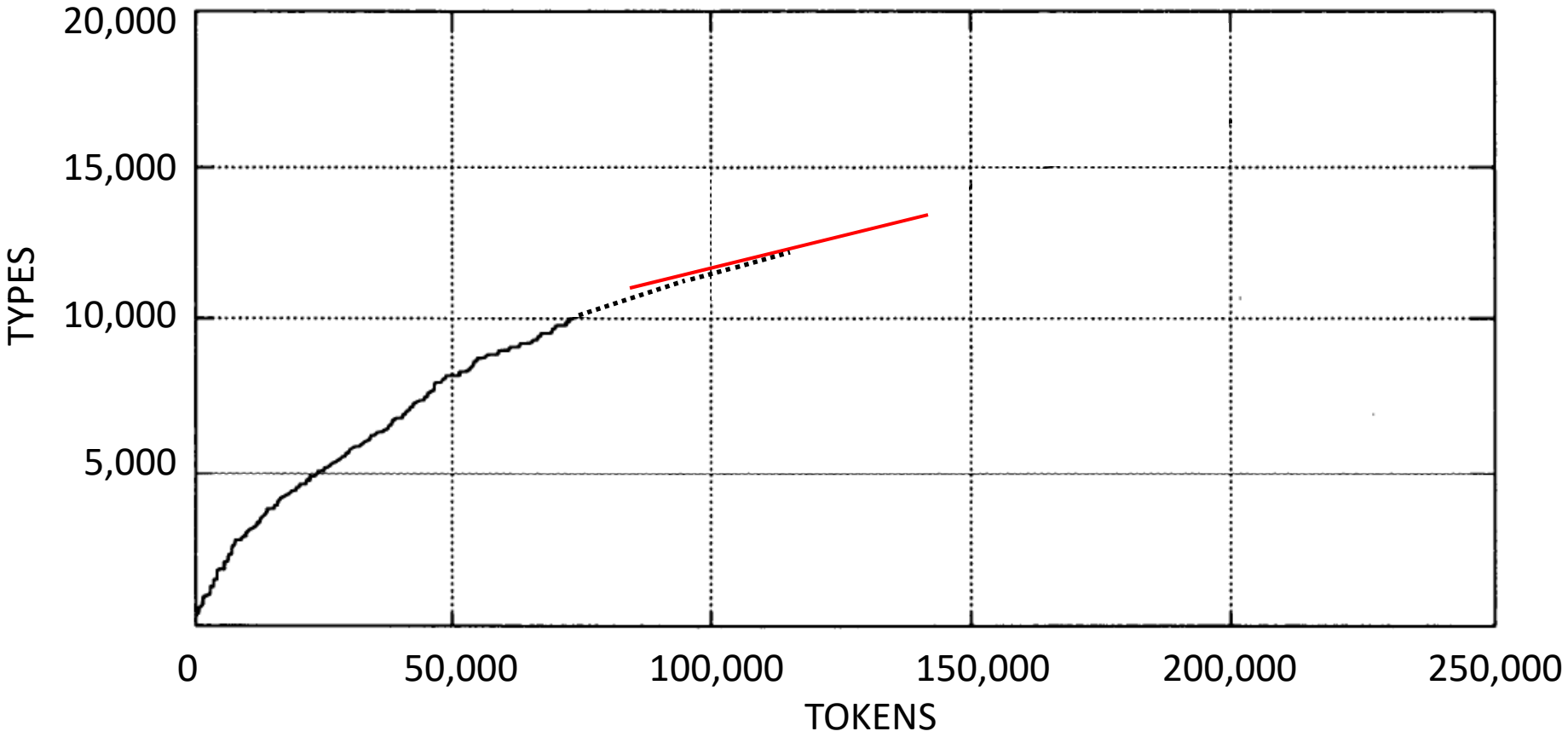

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

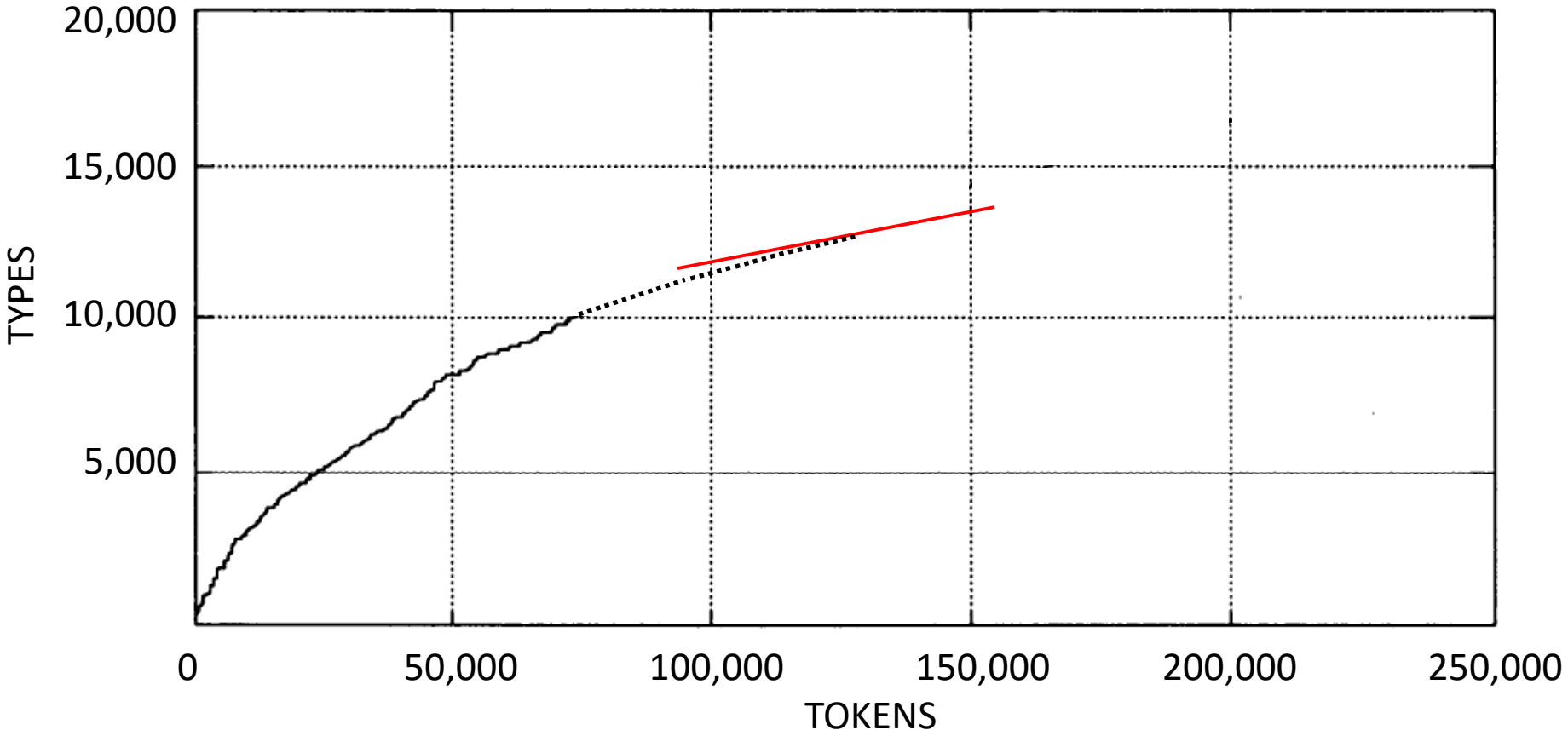

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

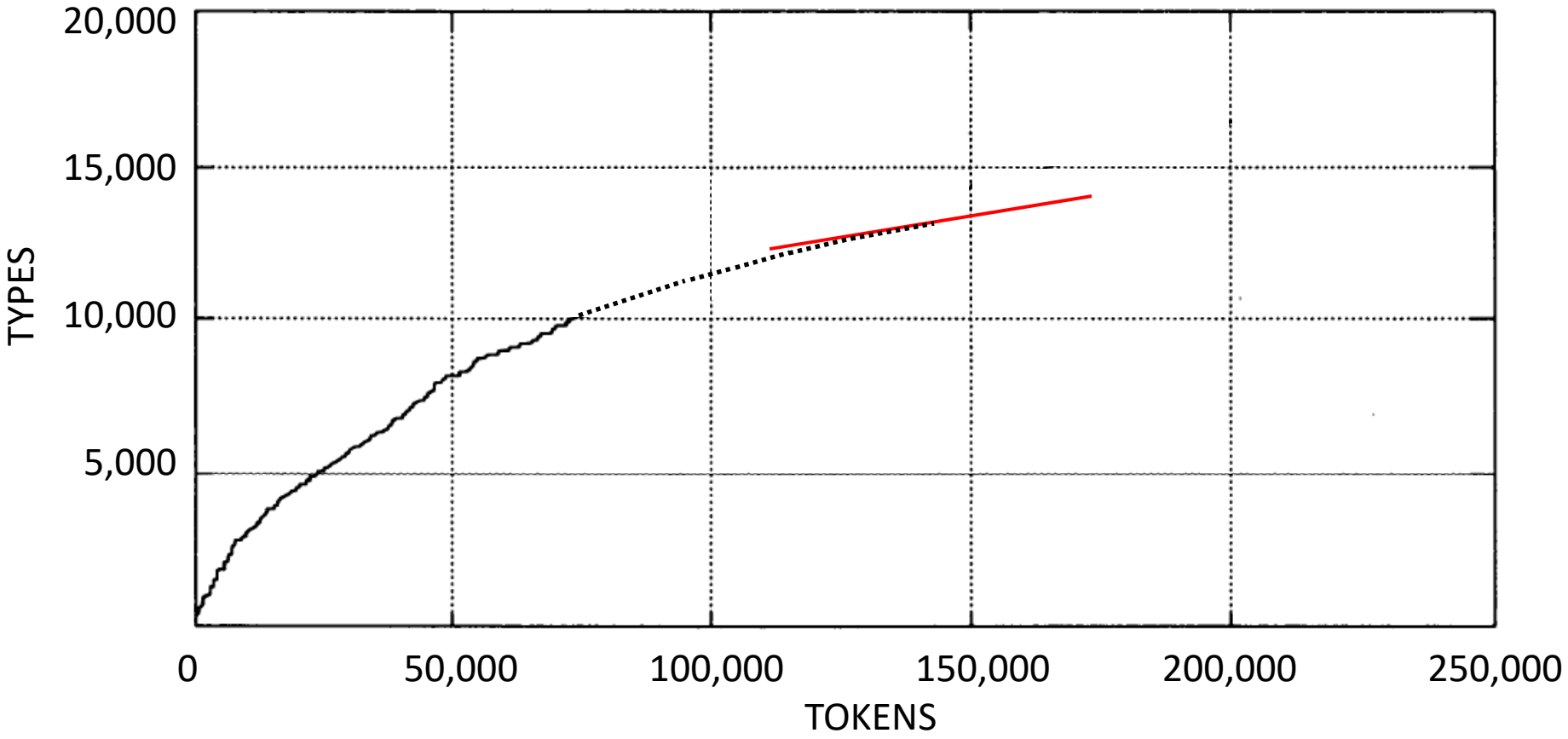

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

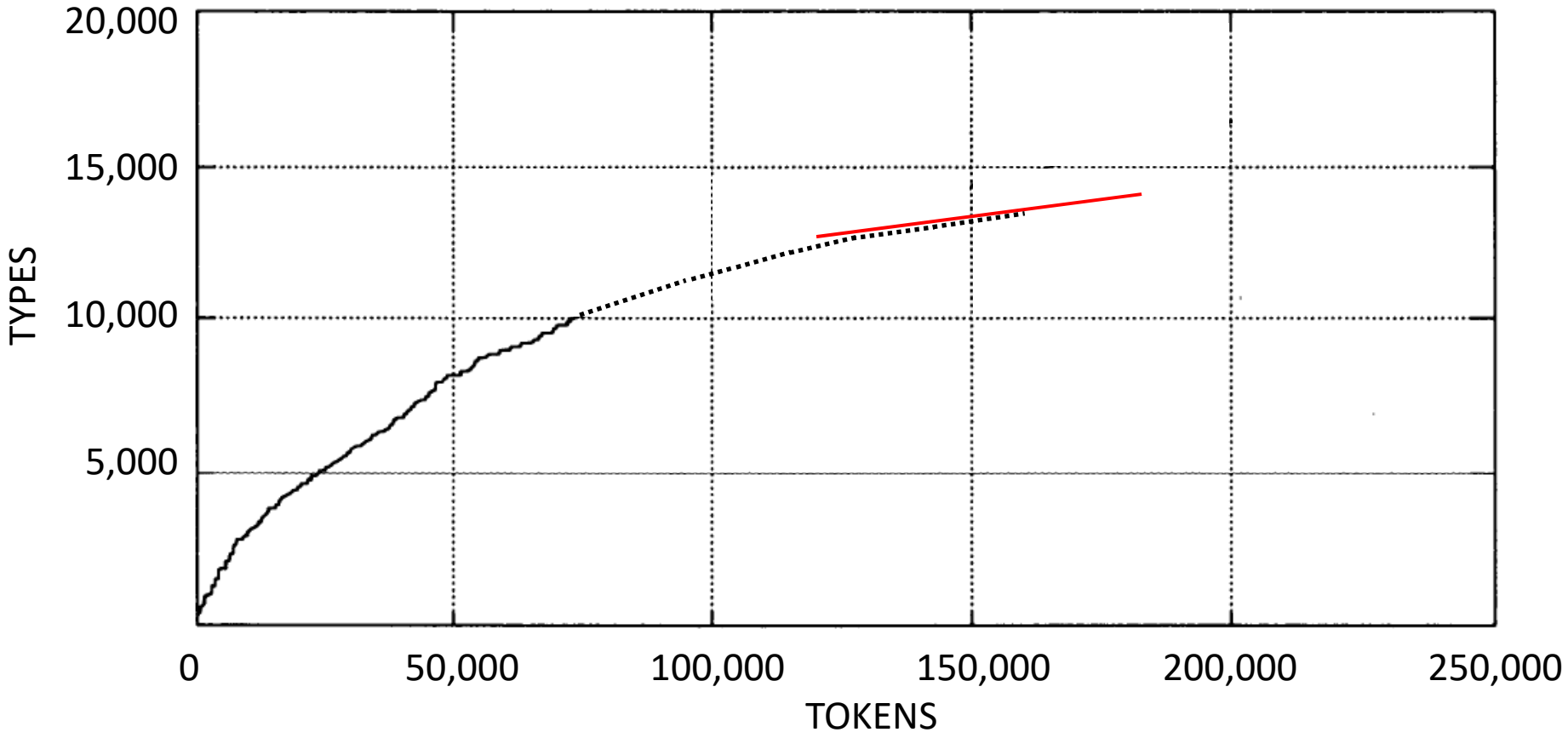

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

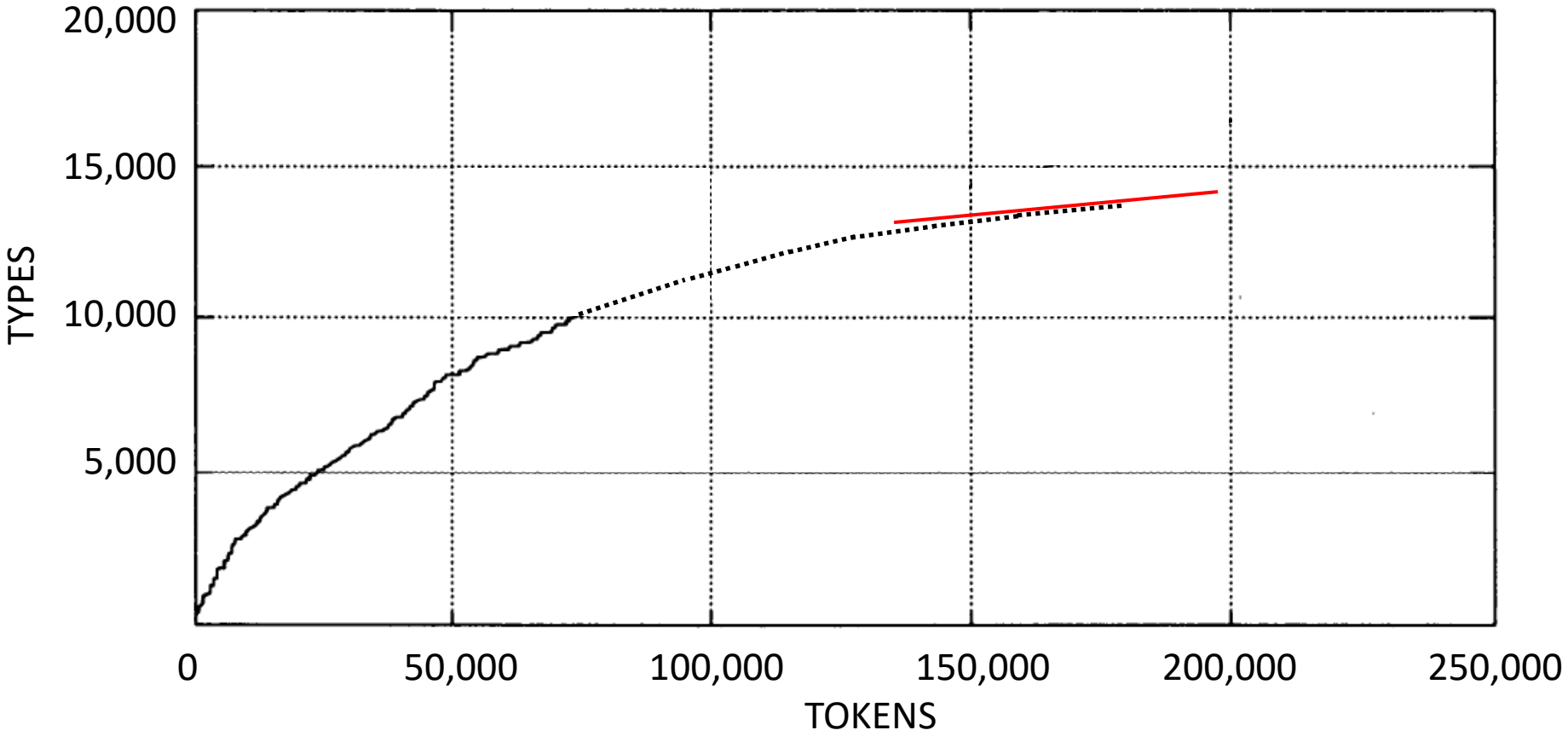

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

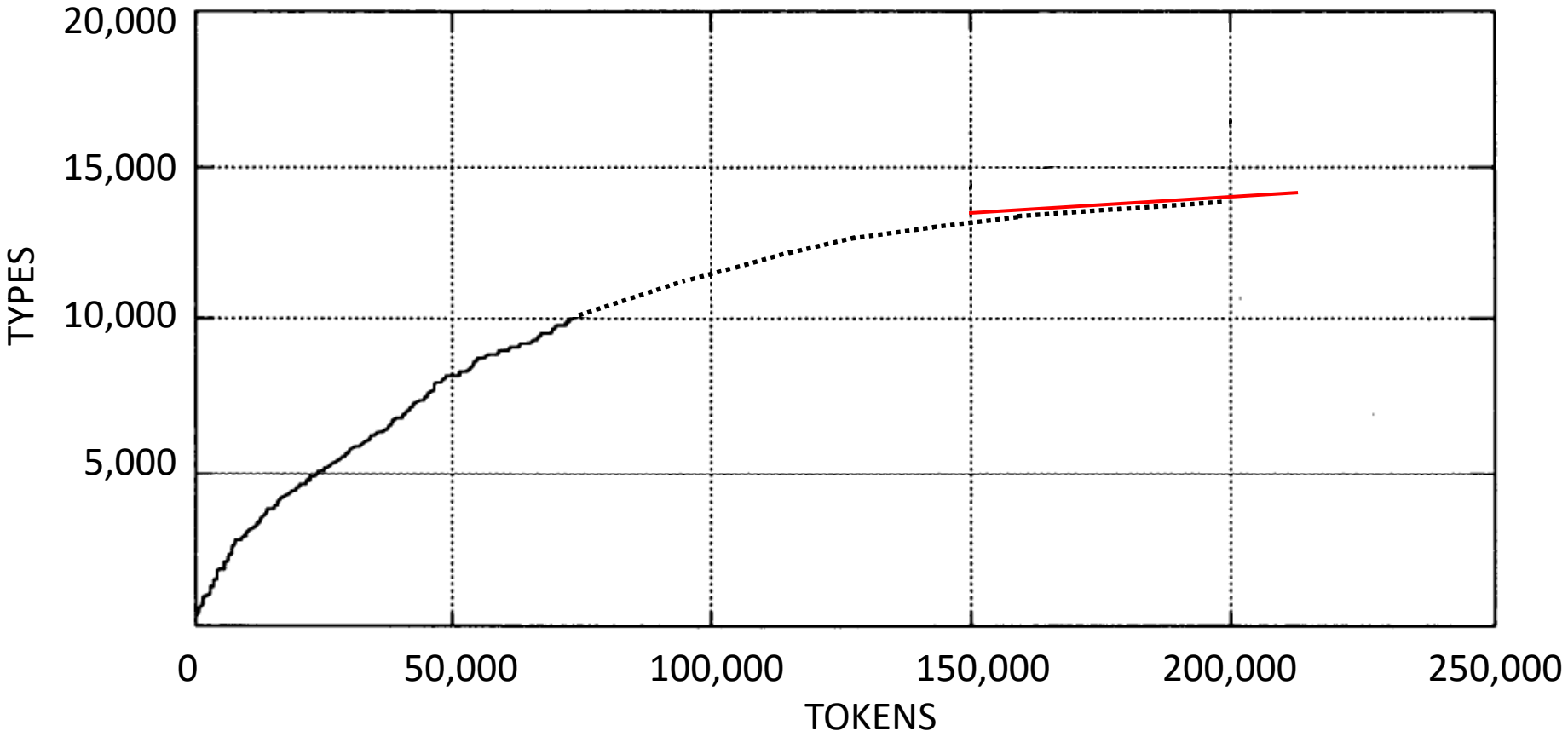

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

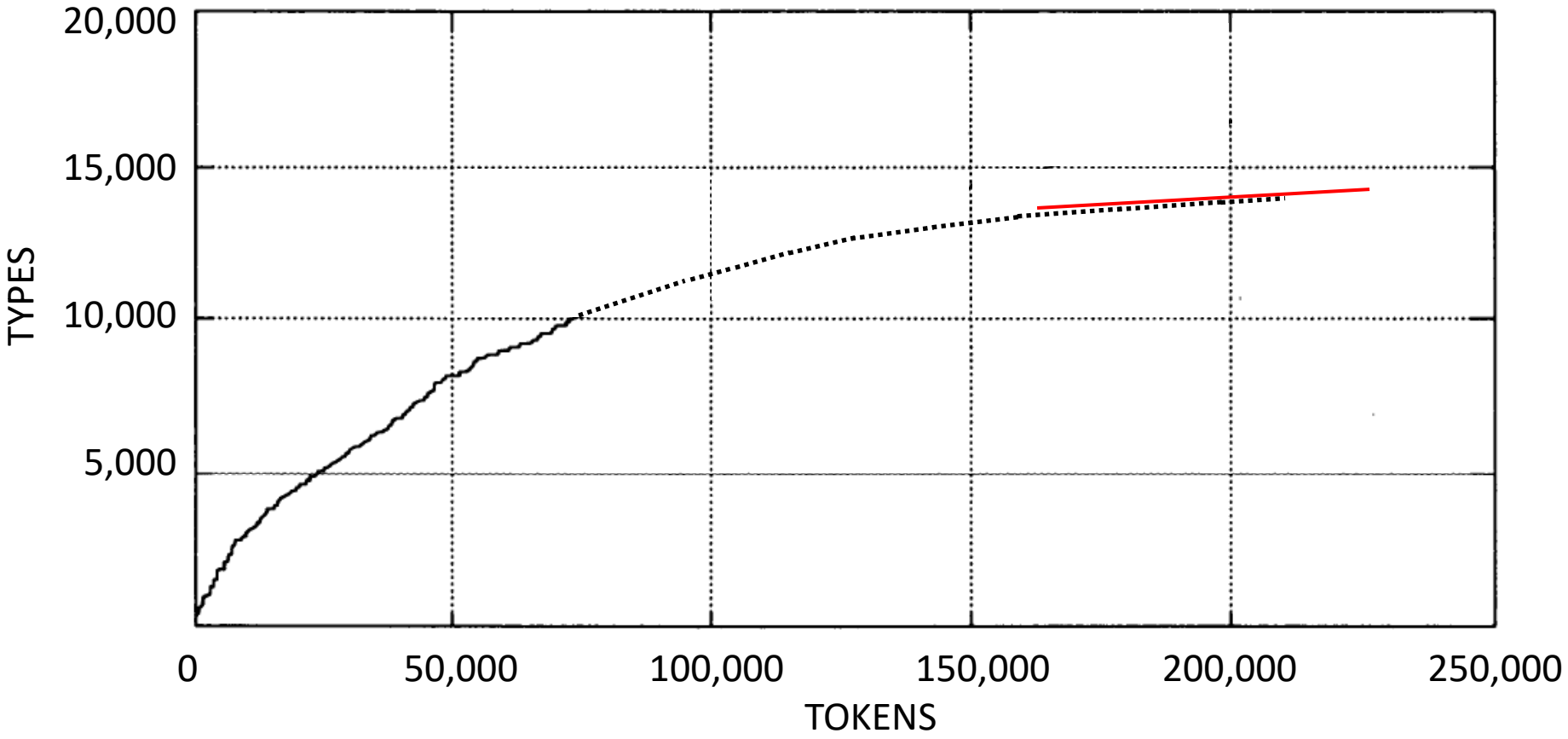

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

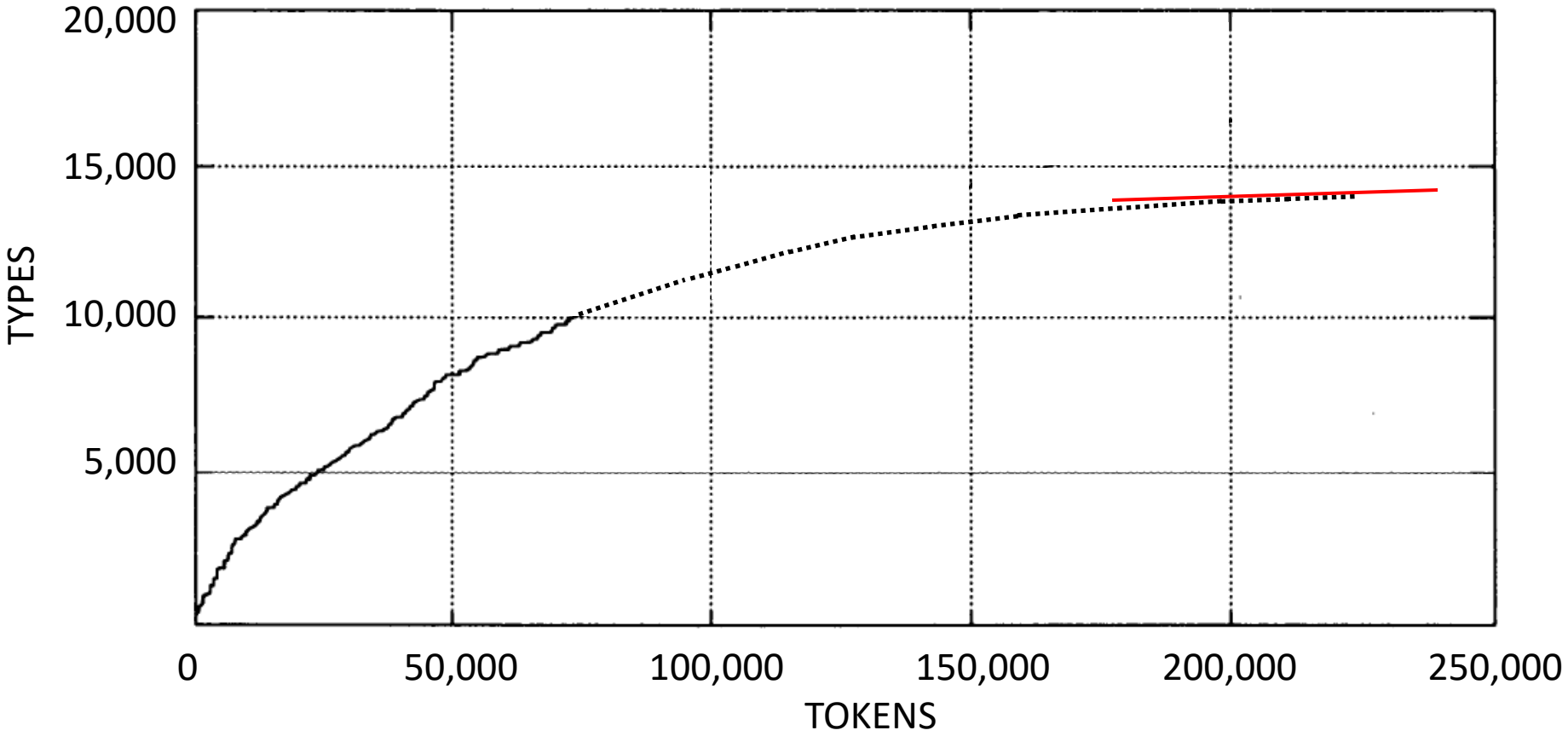

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS


20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS


20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

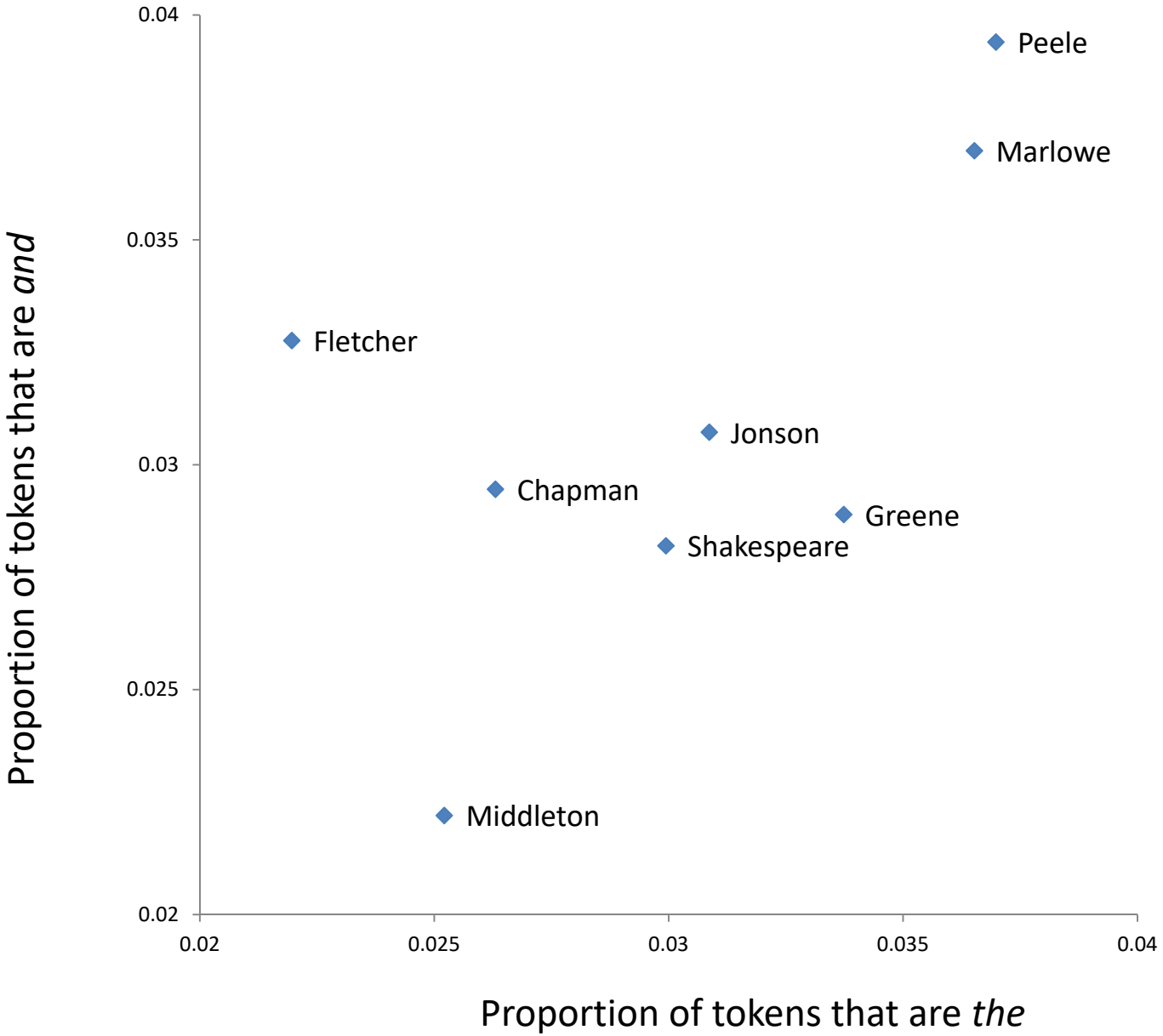

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

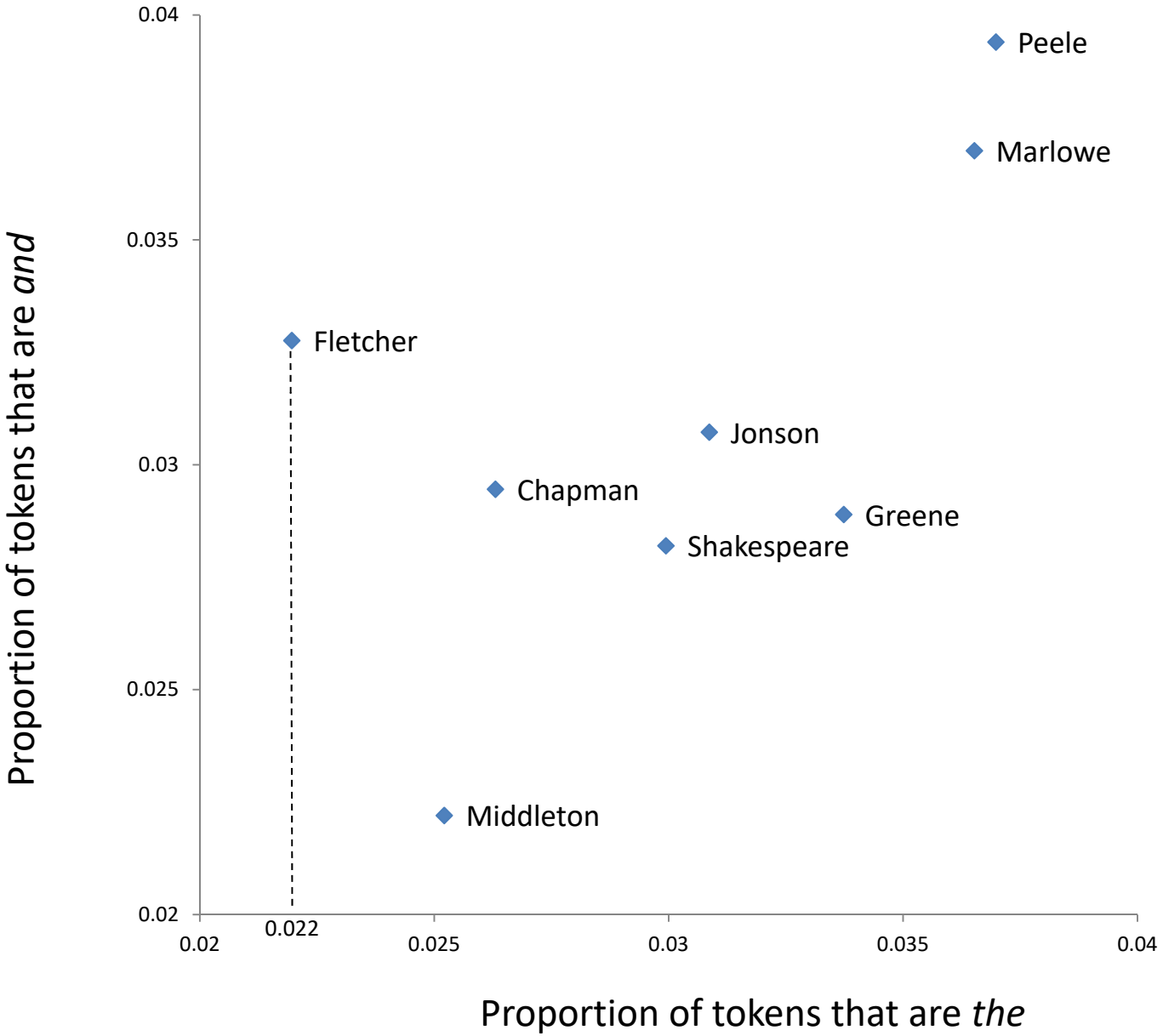

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

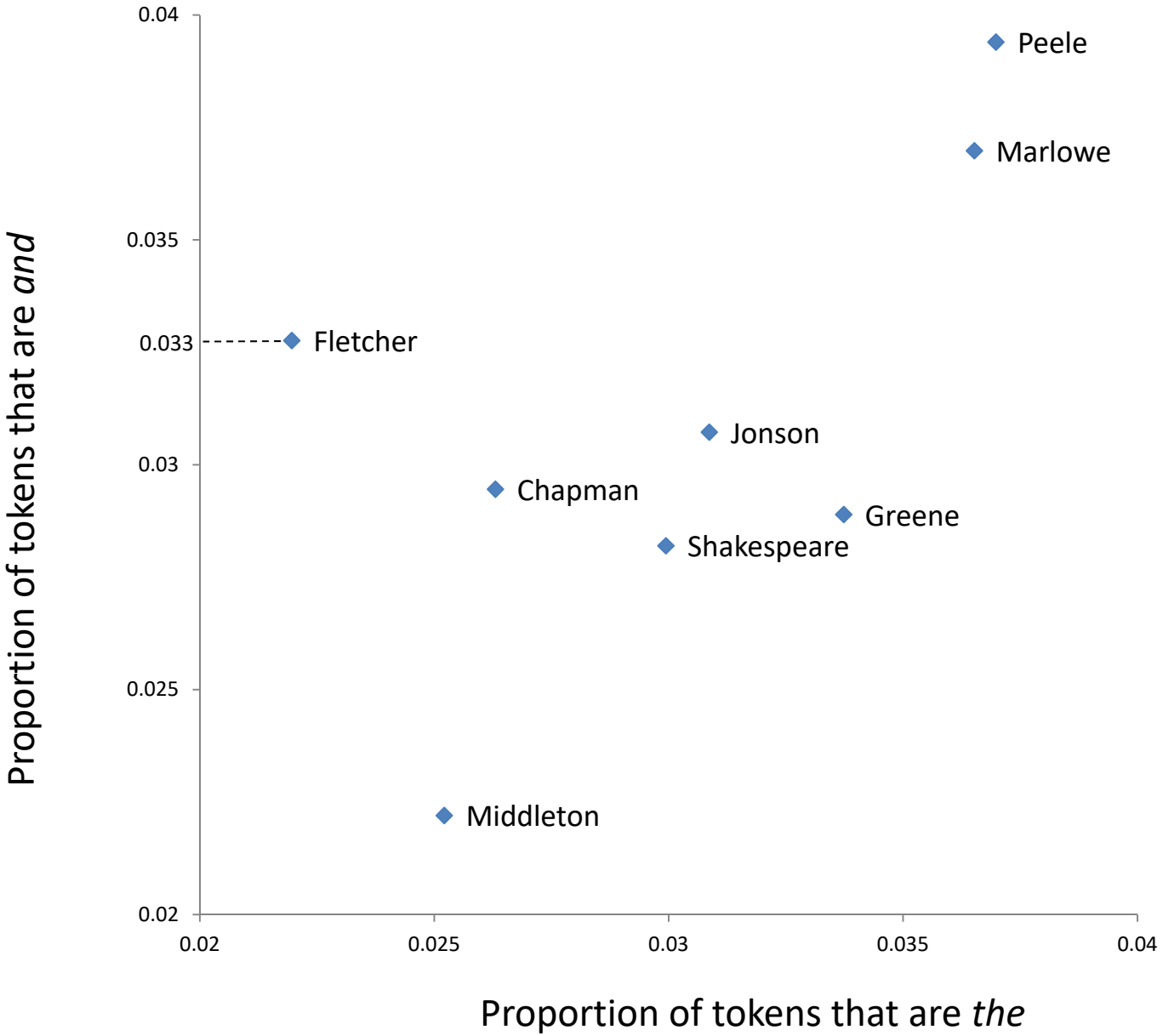

20,000
15,000
10,000
5,000
TYPES
0
50,000
100,000
150,000
200,000
250,000
TOKENS

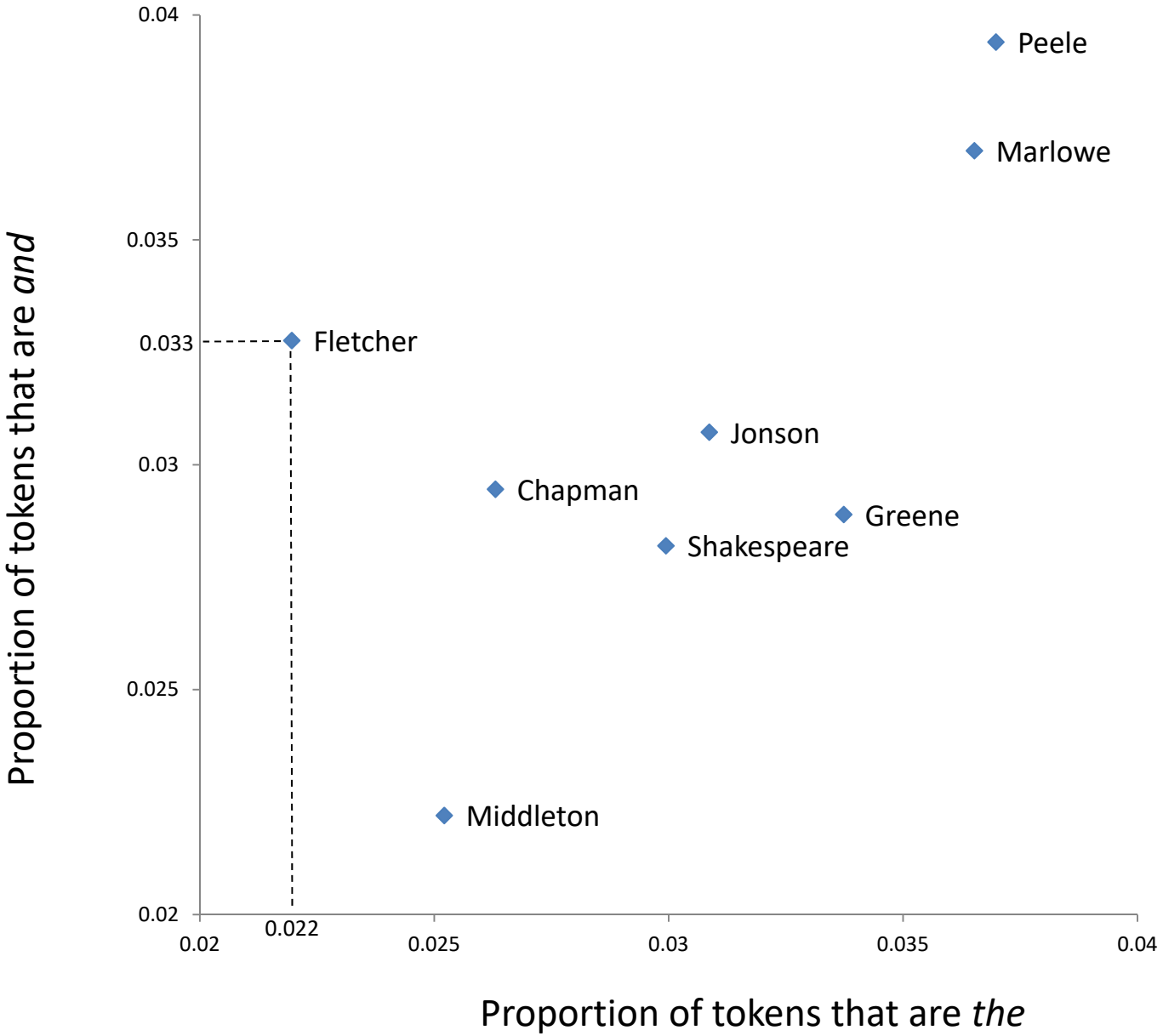
```
punctuation_to_remove = [',', '.', '!', '?', '"', ':', ';', ')', '(', '[', ']', '<', '>', '-', '{', '}', '/', '|']
word_frequencies = {}
file = open ("in.txt", "r")
play = file.read()

for punctuation_mark in punctuation_to_remove:
        play = play.replace(punctuation_mark, "")

for word in play.split():
        word = word.lower()
        if not word in word_frequencies:
                word_frequencies[word] = 1
        else:
                word_frequencies[word] += 1
for w in sorted(word_frequencies, key=word_frequencies.get, reverse=True):
        print (w, word_frequencies[w])

print (len(word_frequencies))
print (len(play.split()))
```

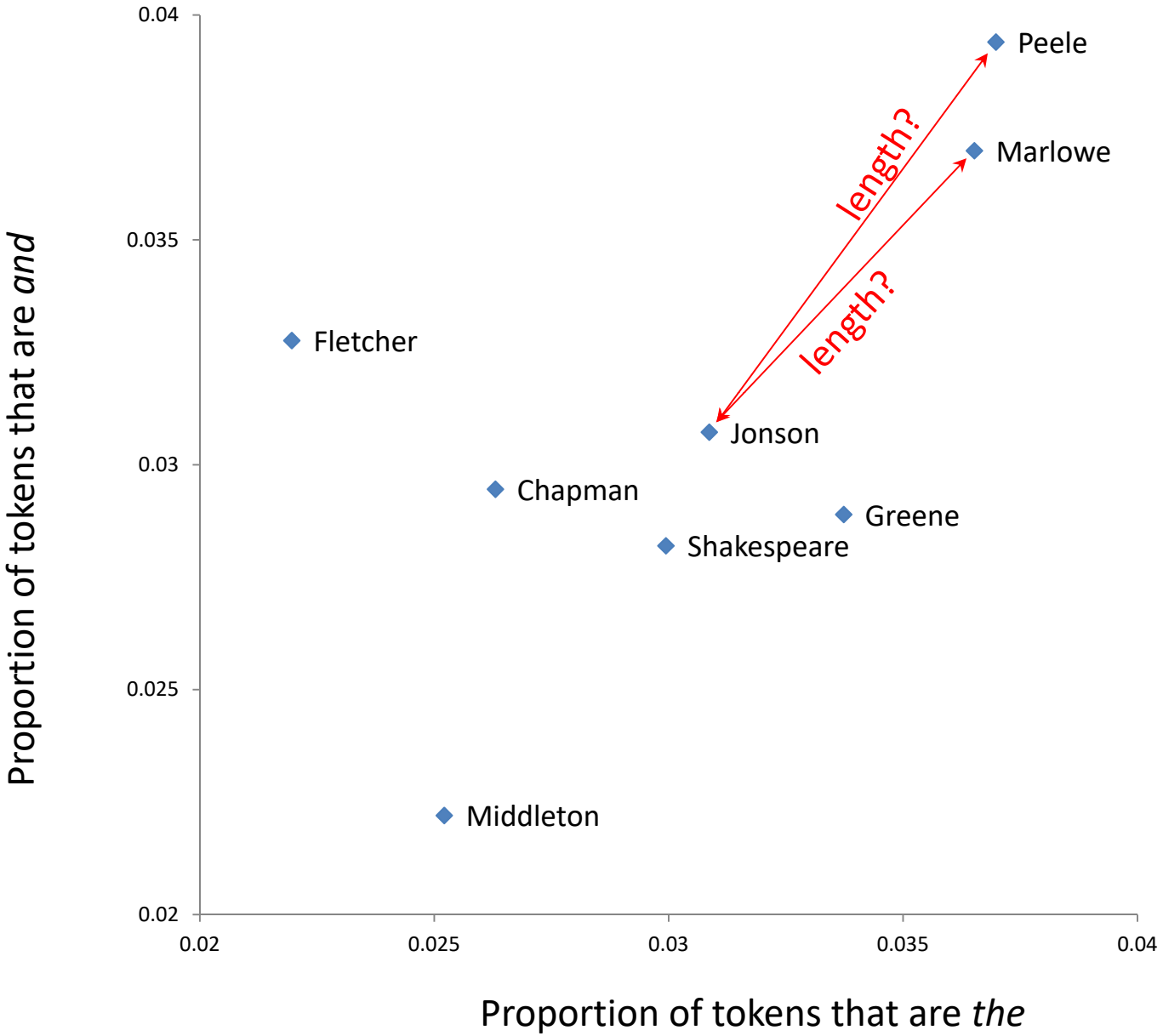
and 6999	but 1823	by 956	thy 654
the 6250	be 1764	lord 936	see 620
to 5365	all 1728	now 910	would 608
i 5314	this 1664	shall 896	which 601
of 4489	me 1646	good 882	ile 600
a 3990	as 1630	are 864	more 599
you 3977	her 1600	their 839	man 596
my 3838	so 1540	at 753	must 573
in 3387	him 1514	or 731	was 560
that 2922	he 1509	well 713	them 556
your 2654	haue 1390	thou 694	one 553
for 2305	will 1341	come 692	how 551
not 2290	what 1175	on 678	we 537
his 2162	no 1138	from 669	make 529
it 2120	sir 1067	they 667	yet 525
with 2051	then 1038	loue 665	know 519
is 2039	if 970	our 656	...

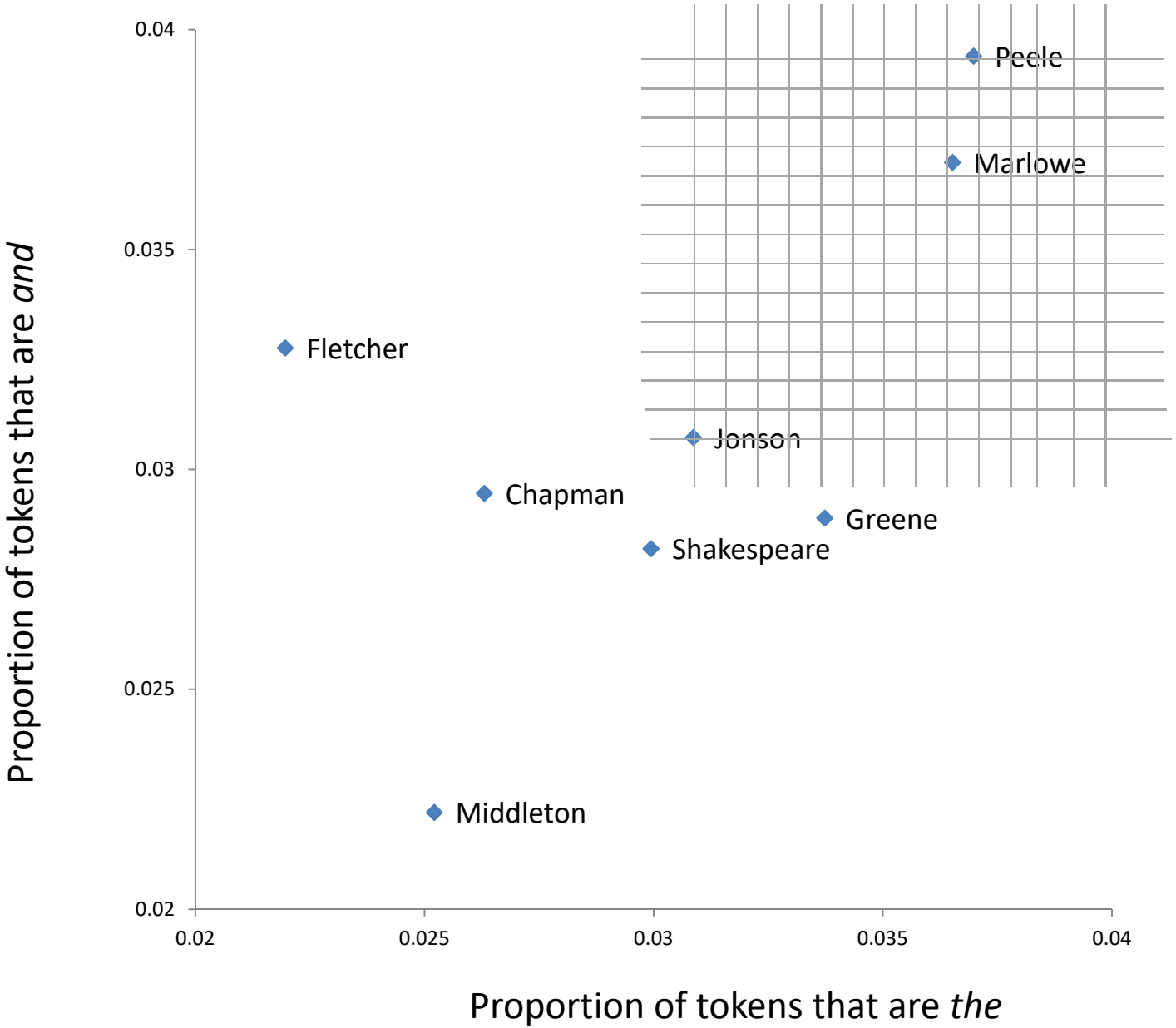

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton
Proportion of tokens that are *and*
Proportion of tokens that are *the*

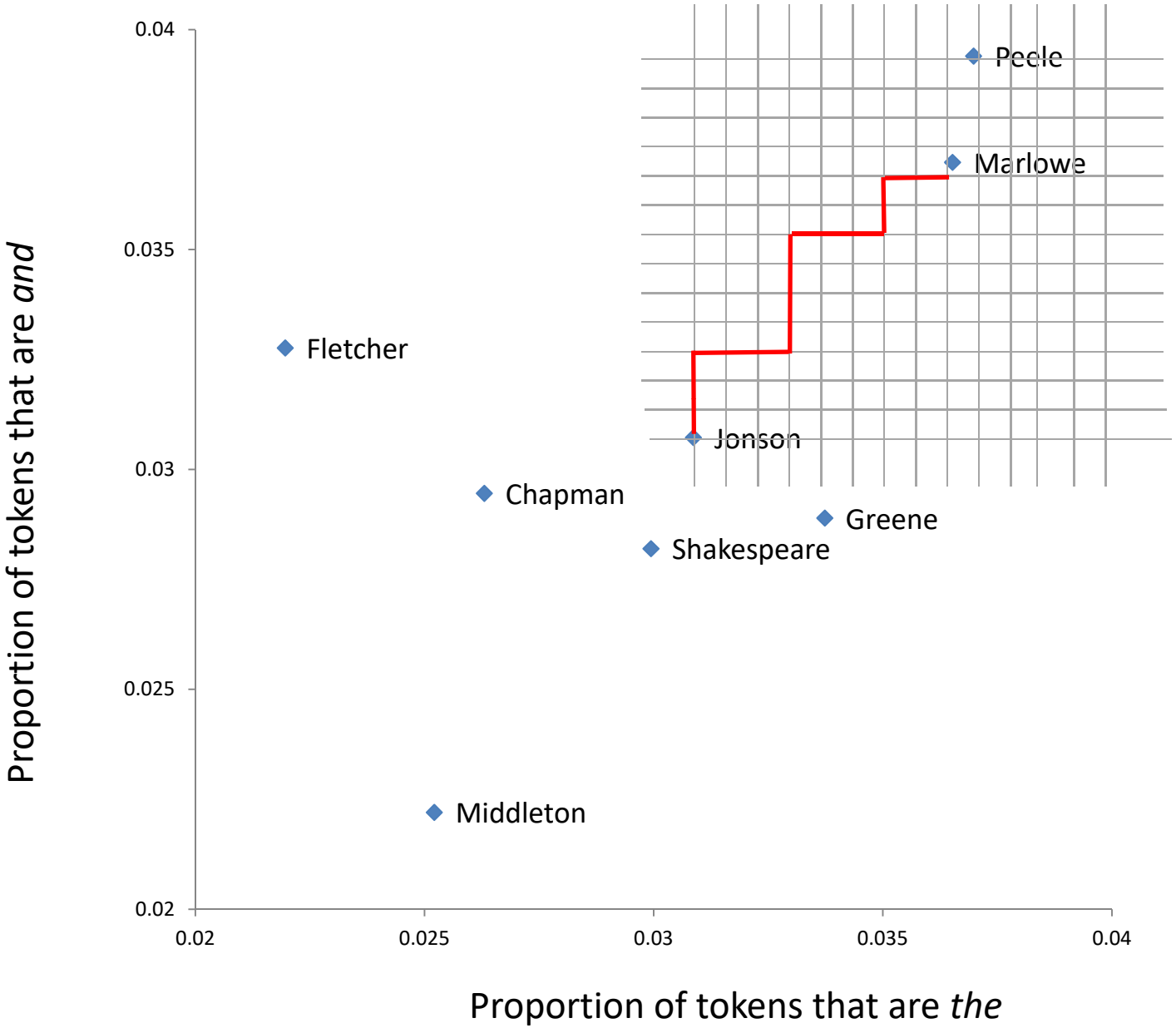

0.04
0.035
0.03
0.025
0.02
Proportion of tokens that are *and*
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton
0.02
0.022
0.025
0.03
0.035
0.04
Proportion of tokens that are *the*

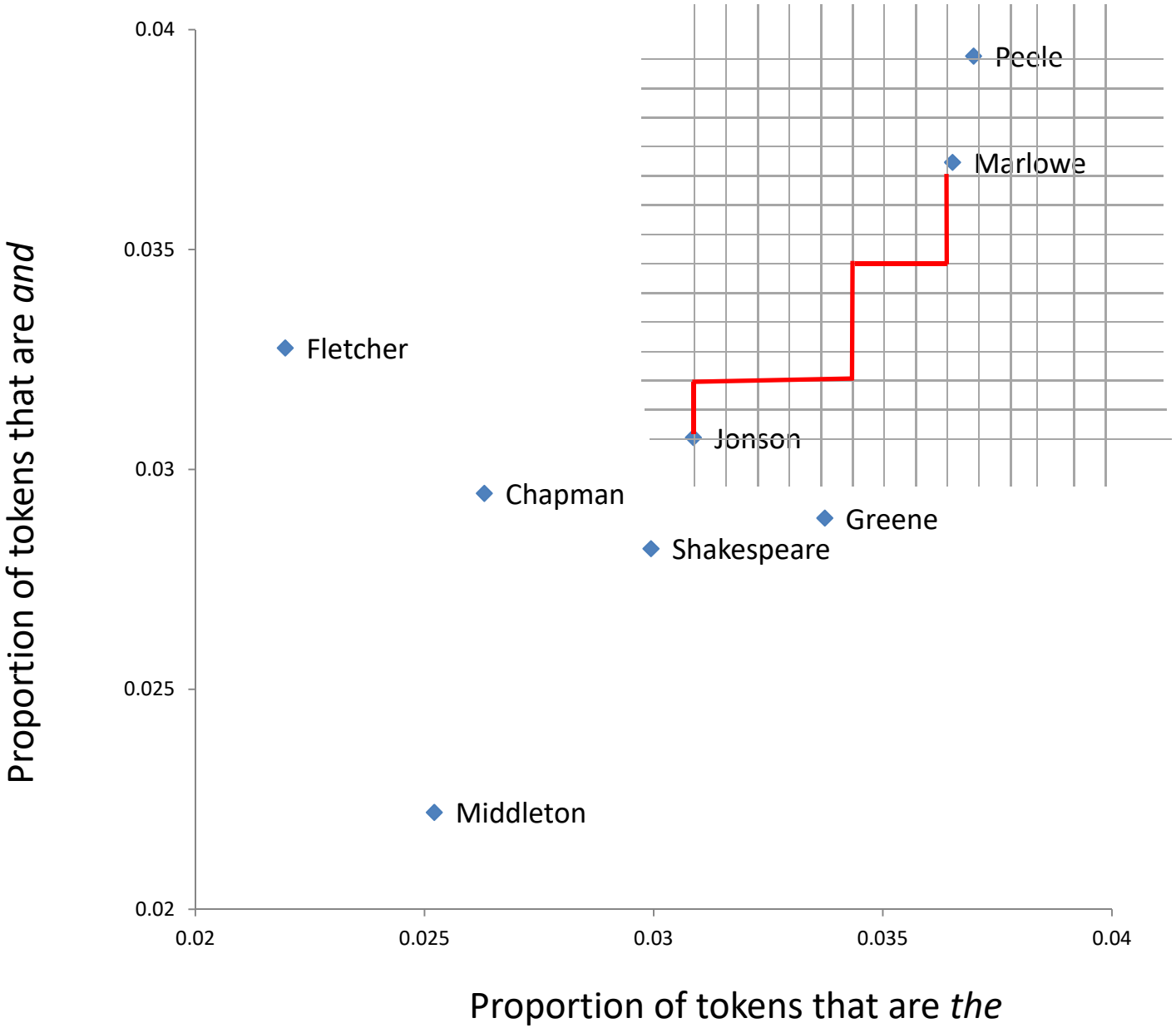

0.04
0.035
0.033
0.03
0.025
0.02
Proportion of tokens that are *and*
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton
0.02
0.025
0.03
0.035
0.04
Proportion of tokens that are *the*

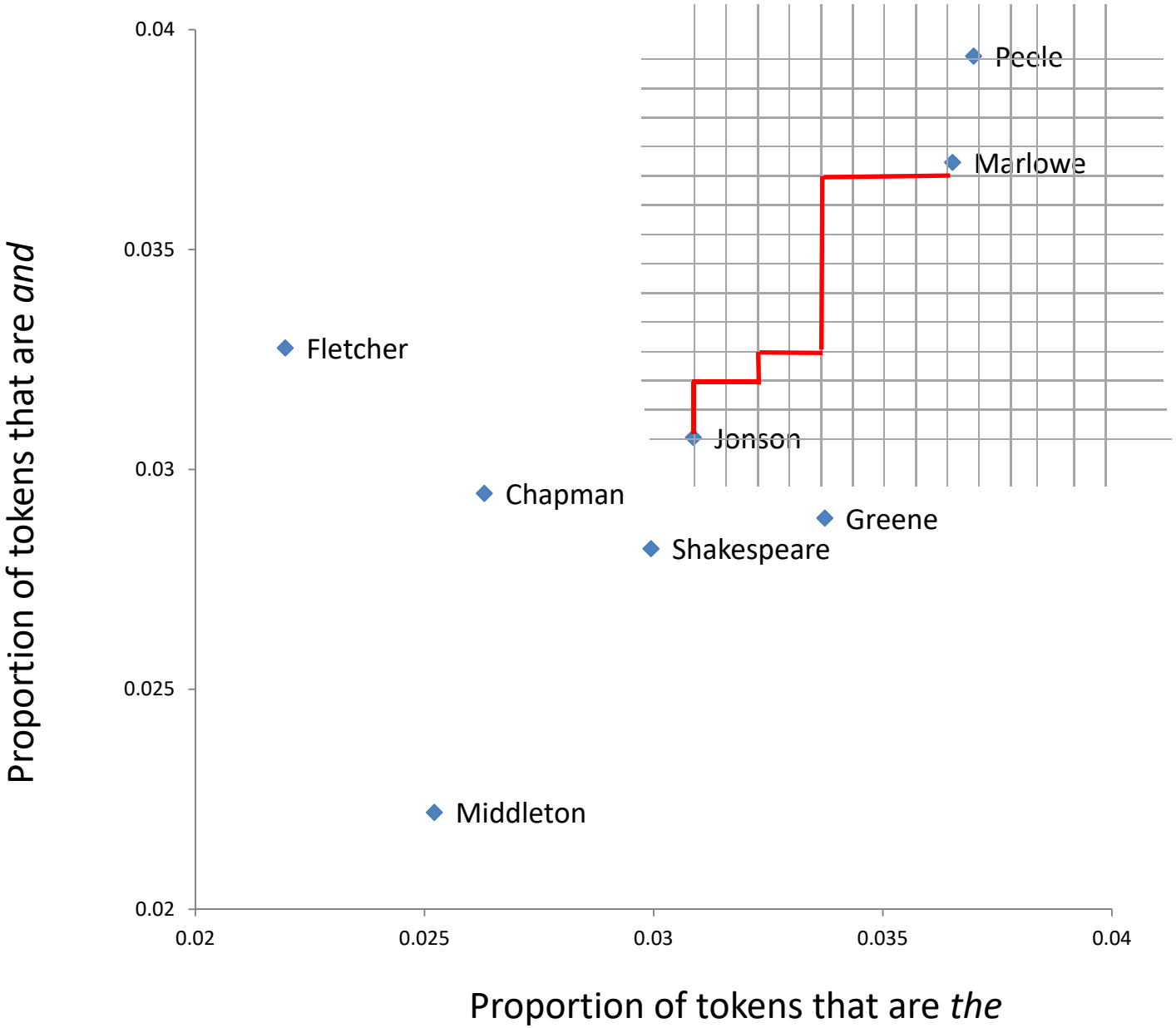

0.04
0.035
0.033
0.03
0.025
0.02
0.02
0.022
0.025
0.03
0.035
0.04
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton
Proportion of tokens that are *and*
Proportion of tokens that are *the*

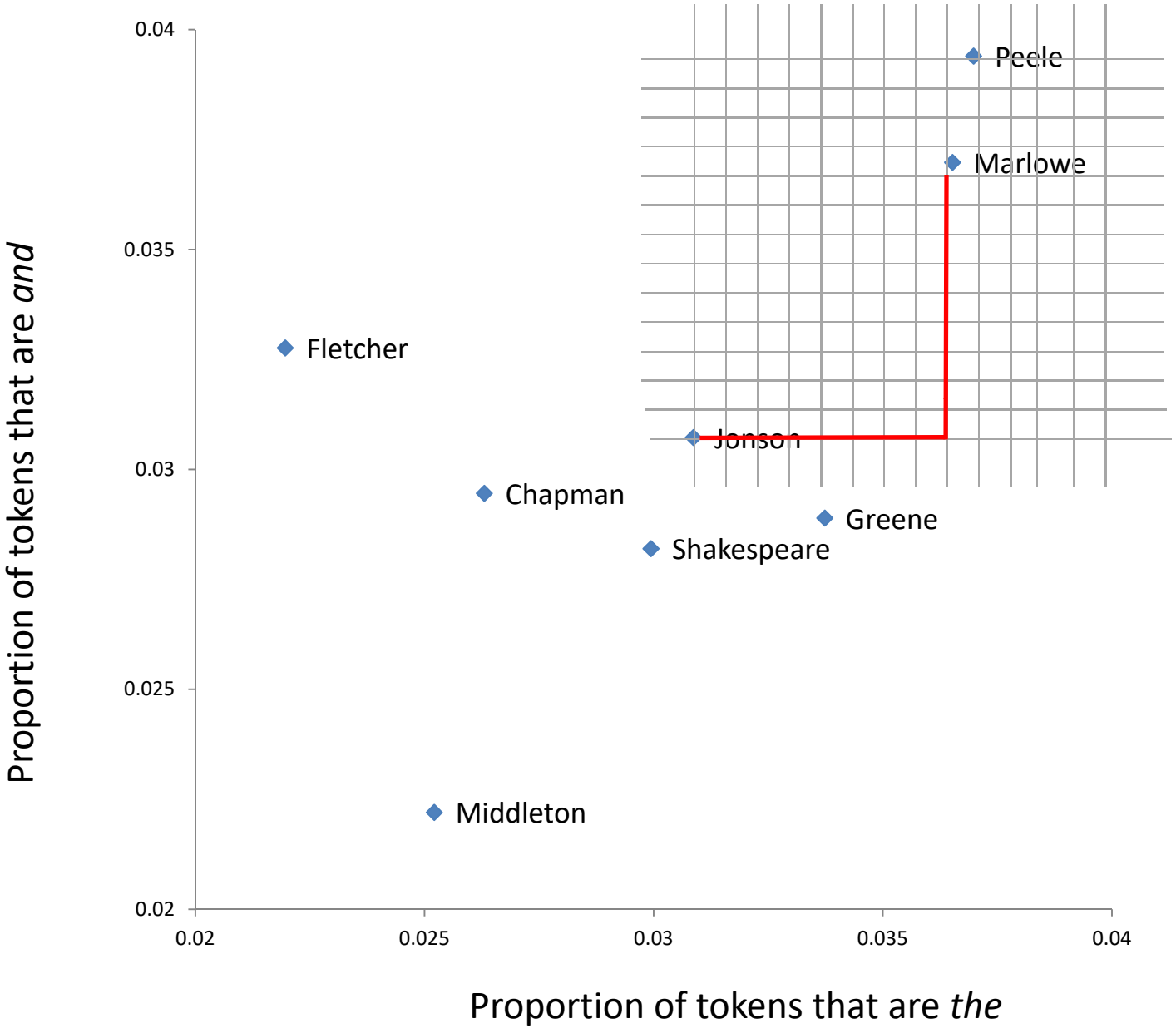

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Proportion of tokens that are *and*
Proportion of tokens that are *the*
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton
length?
length?

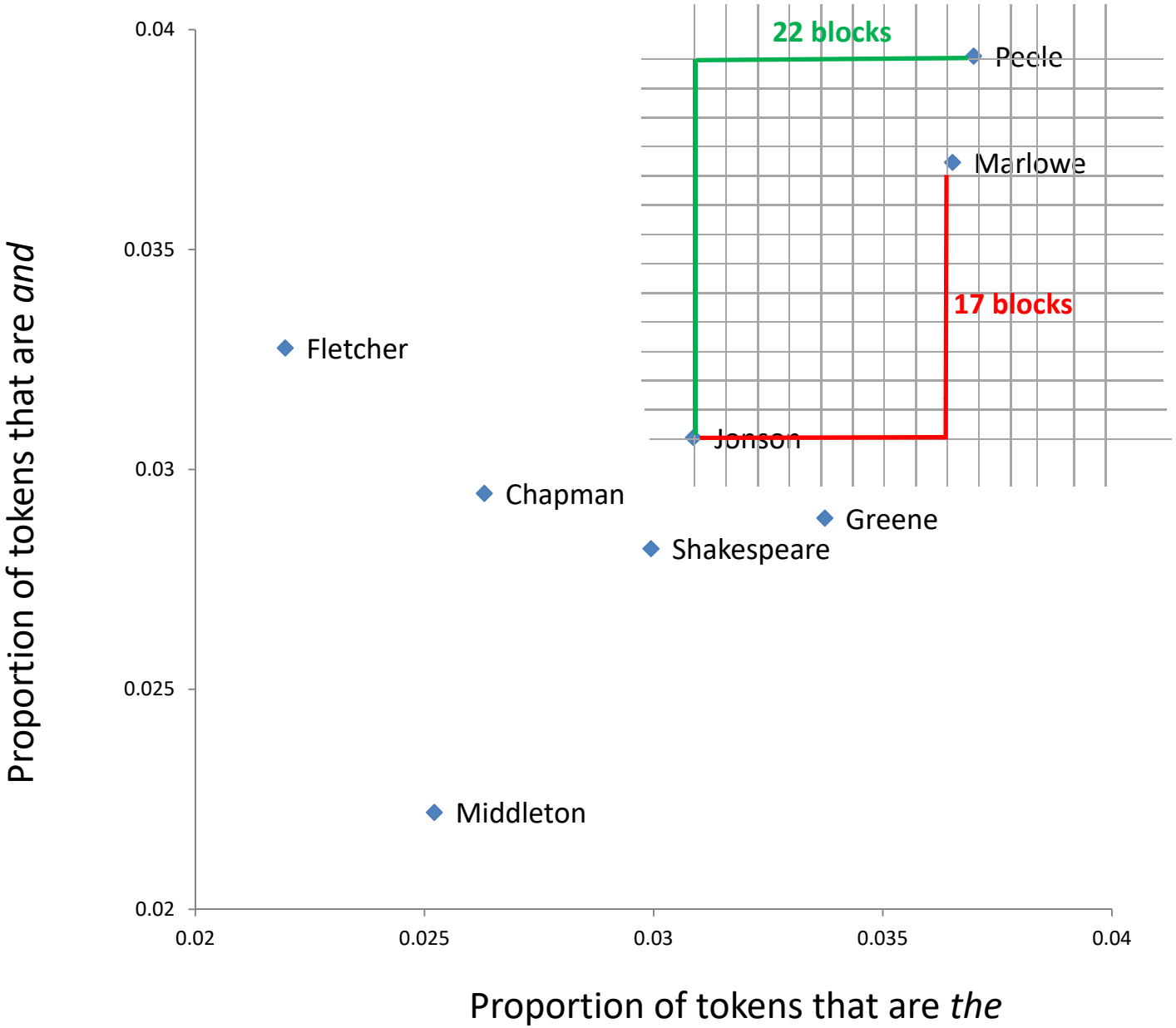

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Proportion of tokens that are *and*
Proportion of tokens that are *the*
Peele
Marlowe
Jonson
Fletcher
Chapman
Greene
Shakespeare
Middleton


0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Proportion of tokens that are *and*
Proportion of tokens that are *the*
Peele
Marlowe
Jonson
Fletcher
Chapman
Greene
Shakespeare
Middleton

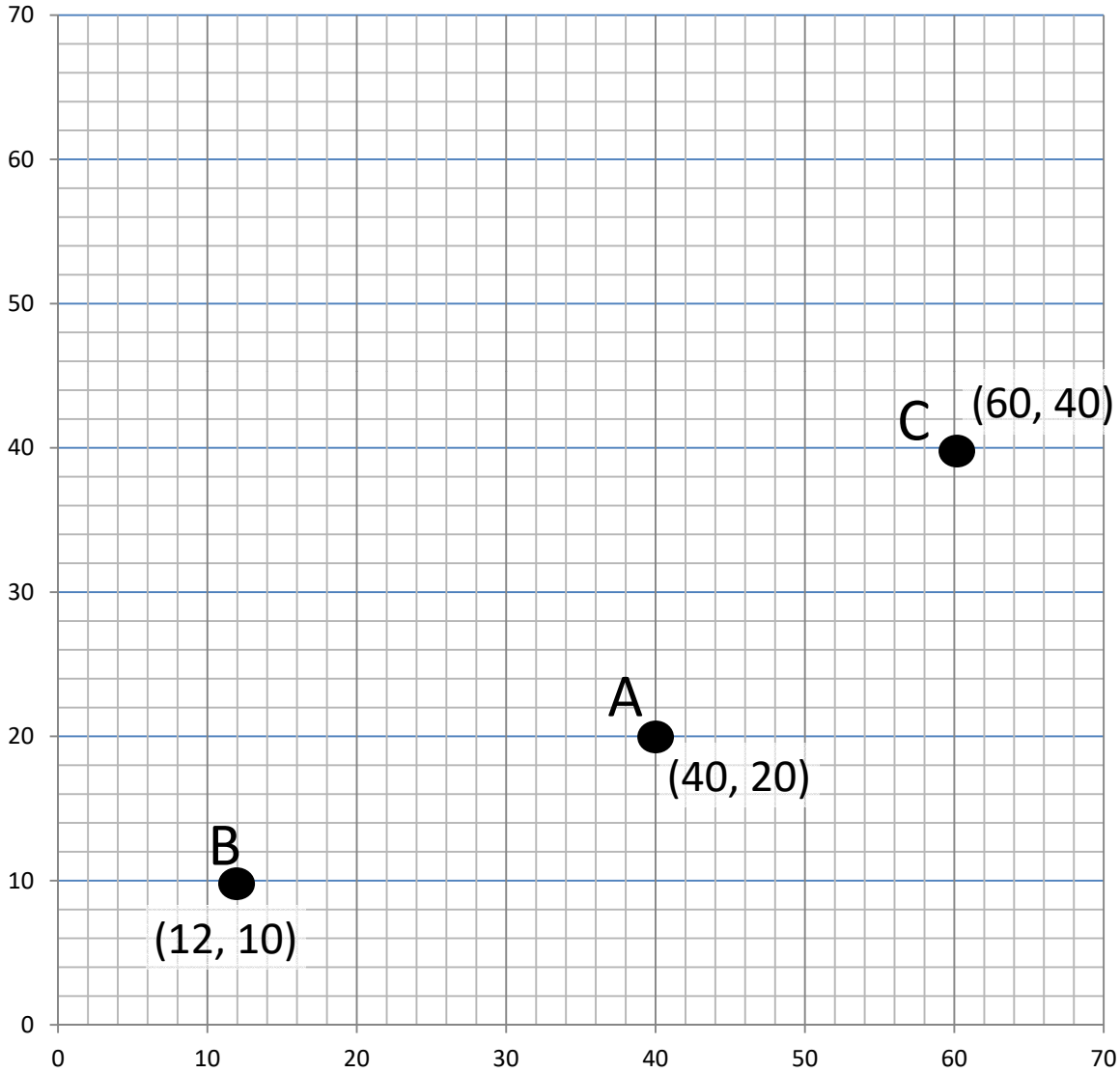

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Proportion of tokens that are *and*
Proportion of tokens that are *the*
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton

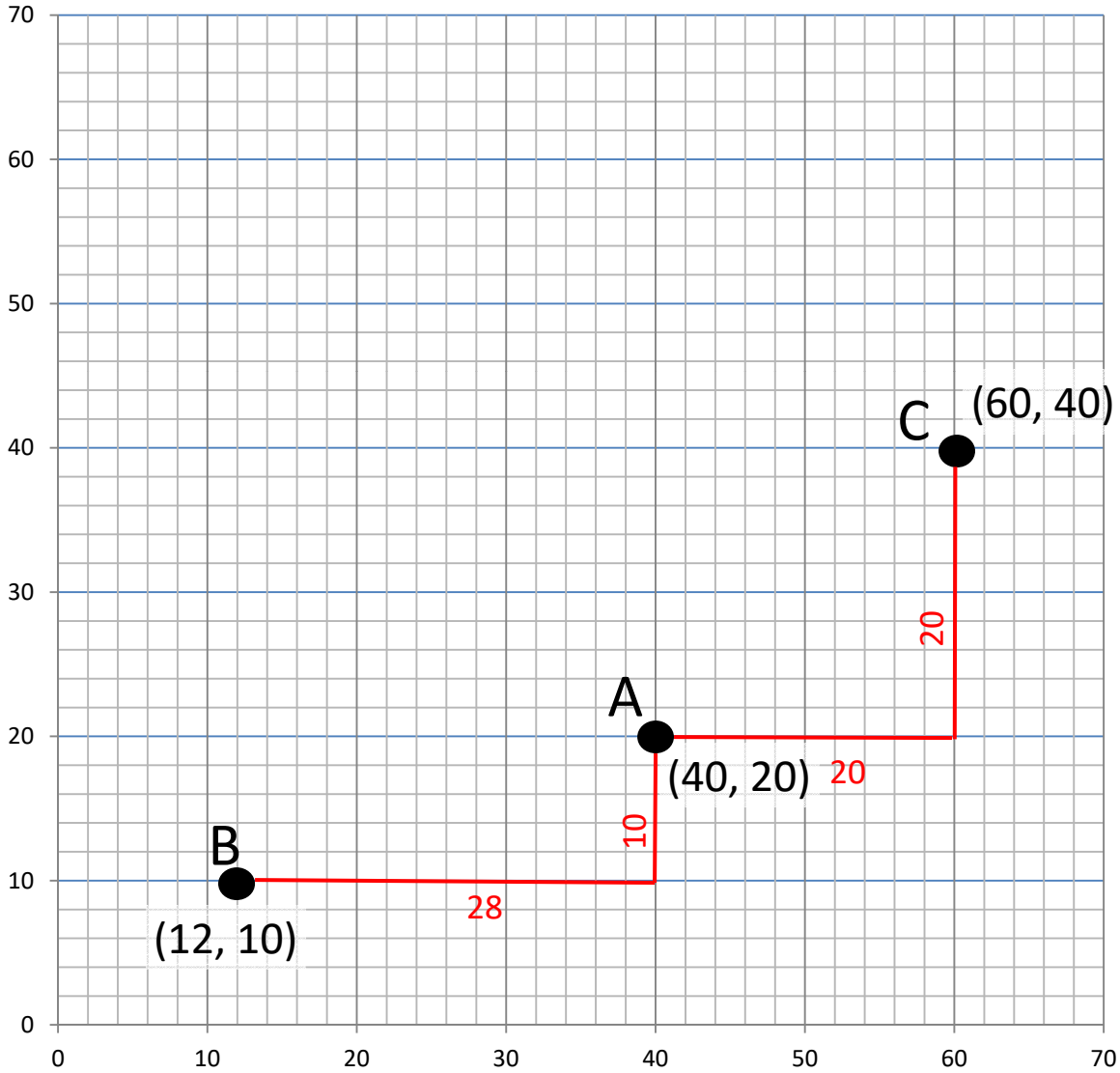

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Proportion of tokens that are *and*
Proportion of tokens that are *the*
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton

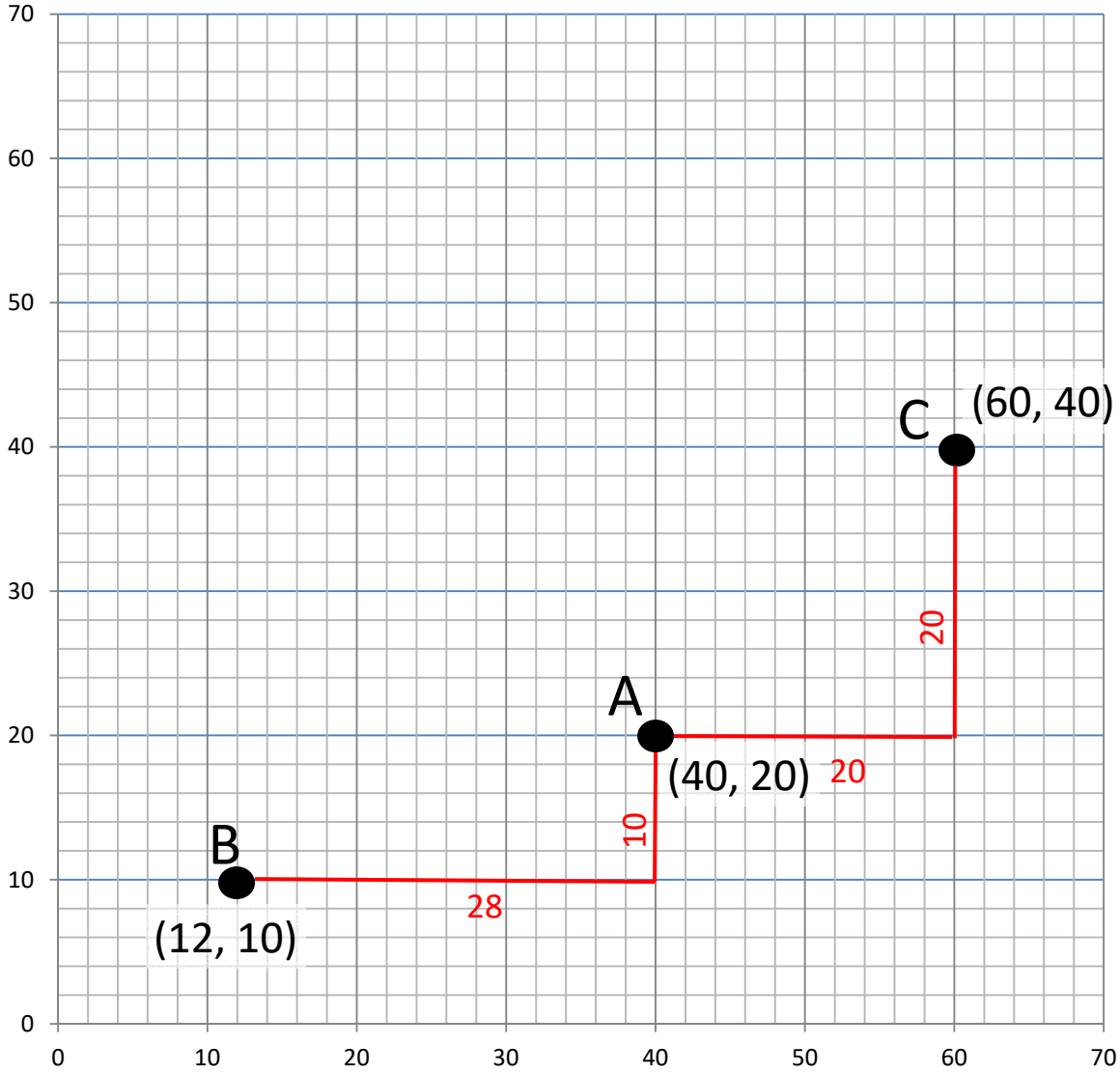

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Proportion of tokens that are *and*
Proportion of tokens that are *the*
Peele
Marlowe
Jonson
Fletcher
Chapman
Greene
Shakespeare
Middleton


22 blocks
17 blocks
Peele
Marlowe
Jonson
Fletcher
Chapman
Greene
Shakespeare
Middleton
0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Proportion of tokens that are *and*
Proportion of tokens that are *the*

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Proportion of tokens that are *and*
Proportion of tokens that are *the*
Fletcher
Chapman
Middleton
Shakespeare
Greene
Peele
Marlowe
Jonson
22 blocks
15.8 blocks
12 blocks
17 blocks


70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
C
(60, 40)
A
(40, 20)
B
(12, 10)


70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
C
(60, 40)
A
(40, 20)
B
(12, 10)
20
20
10
28

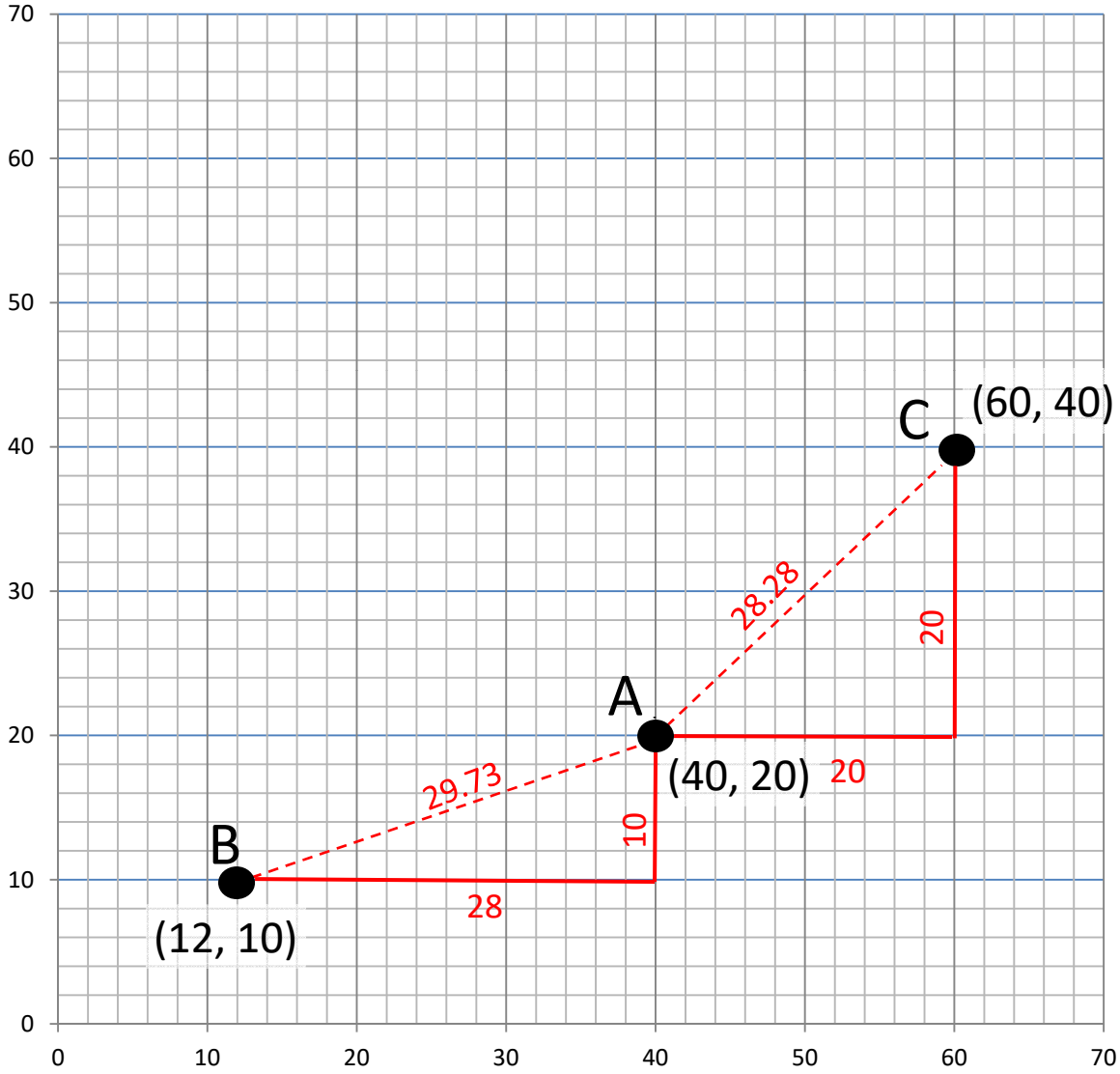


By Manhattan:

A-to-B = 28 + 10 = 38

A-to-C = 20 + 20 = 40

So B is the nearer to A



By Manhattan:

A-to-B = 28 + 10 = 38

A-to-C = 20 + 20 = 40

So B is the nearer to A

By Euclidean:

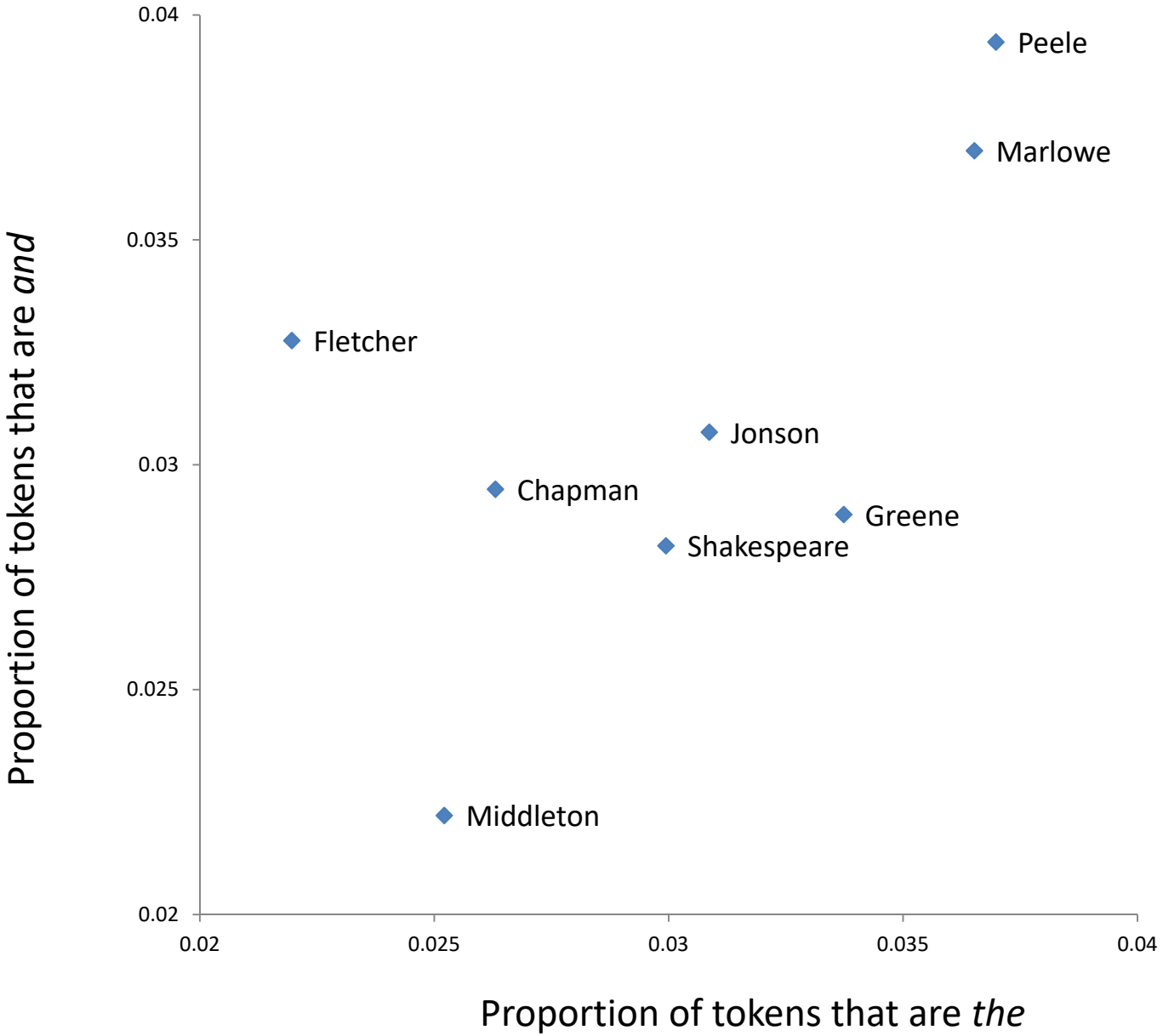
A-to-B = $\sqrt{28^2 + 10^2}$

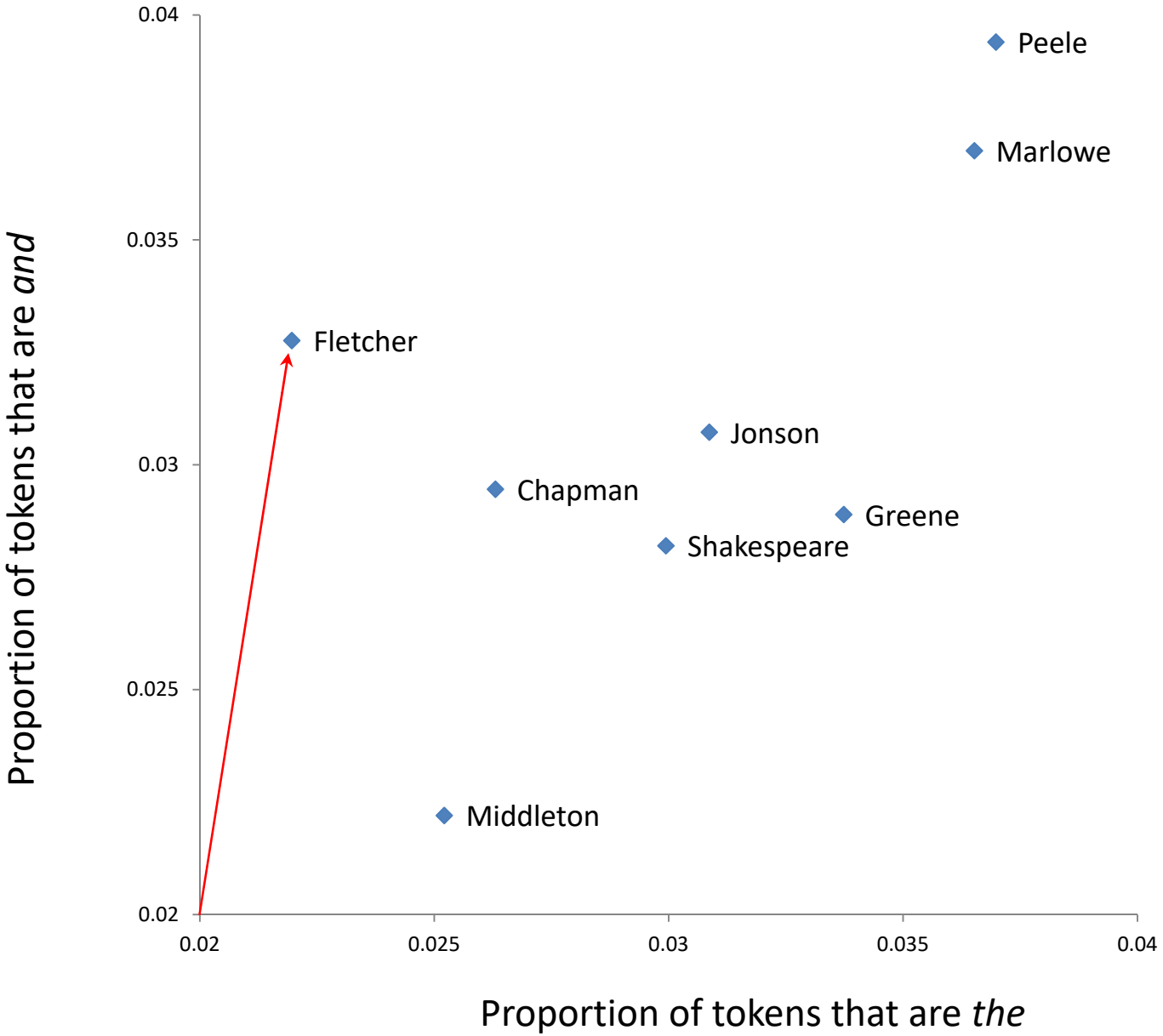
= 29.73

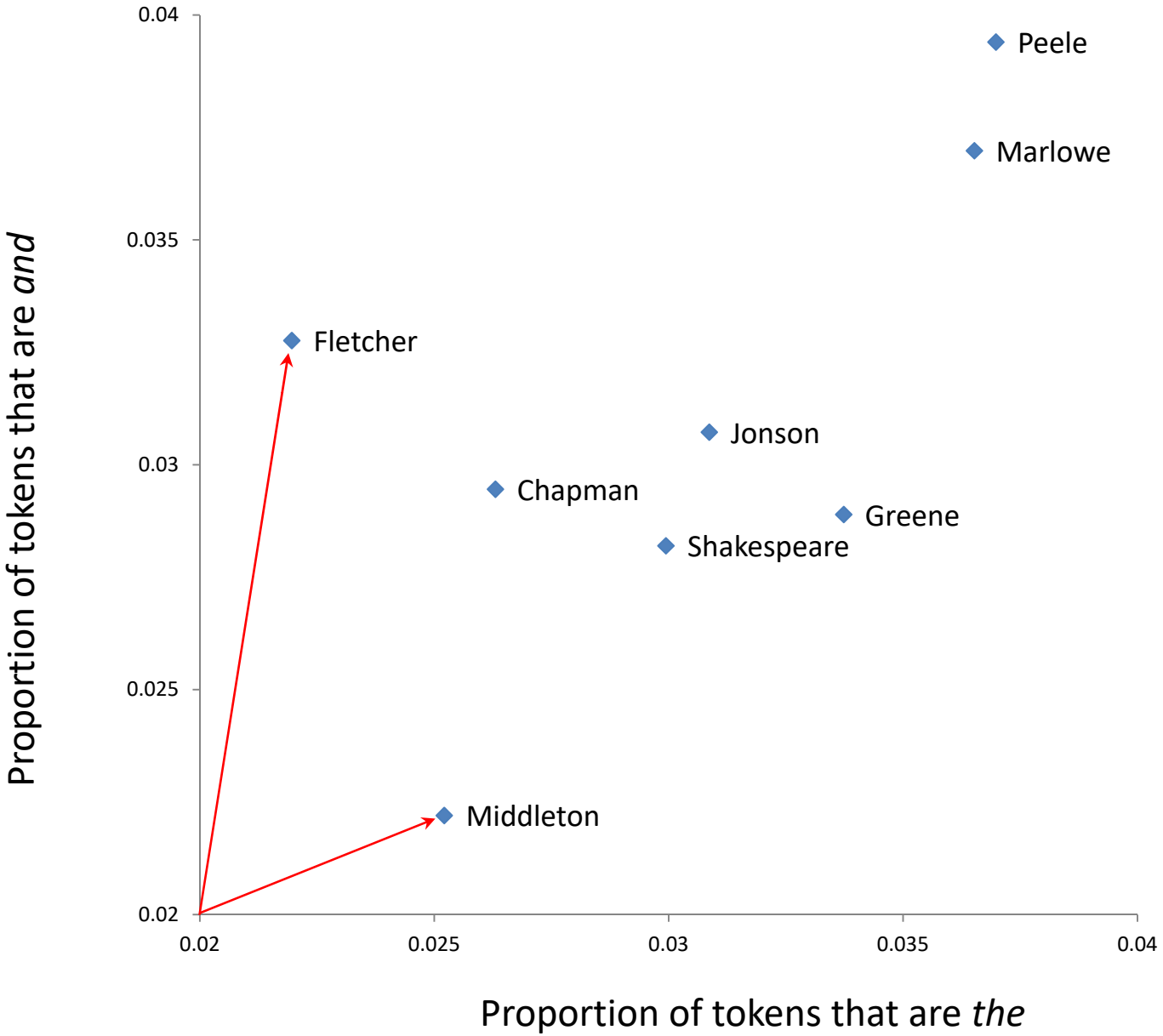
A-to-C = $\sqrt{20^2 + 20^2}$

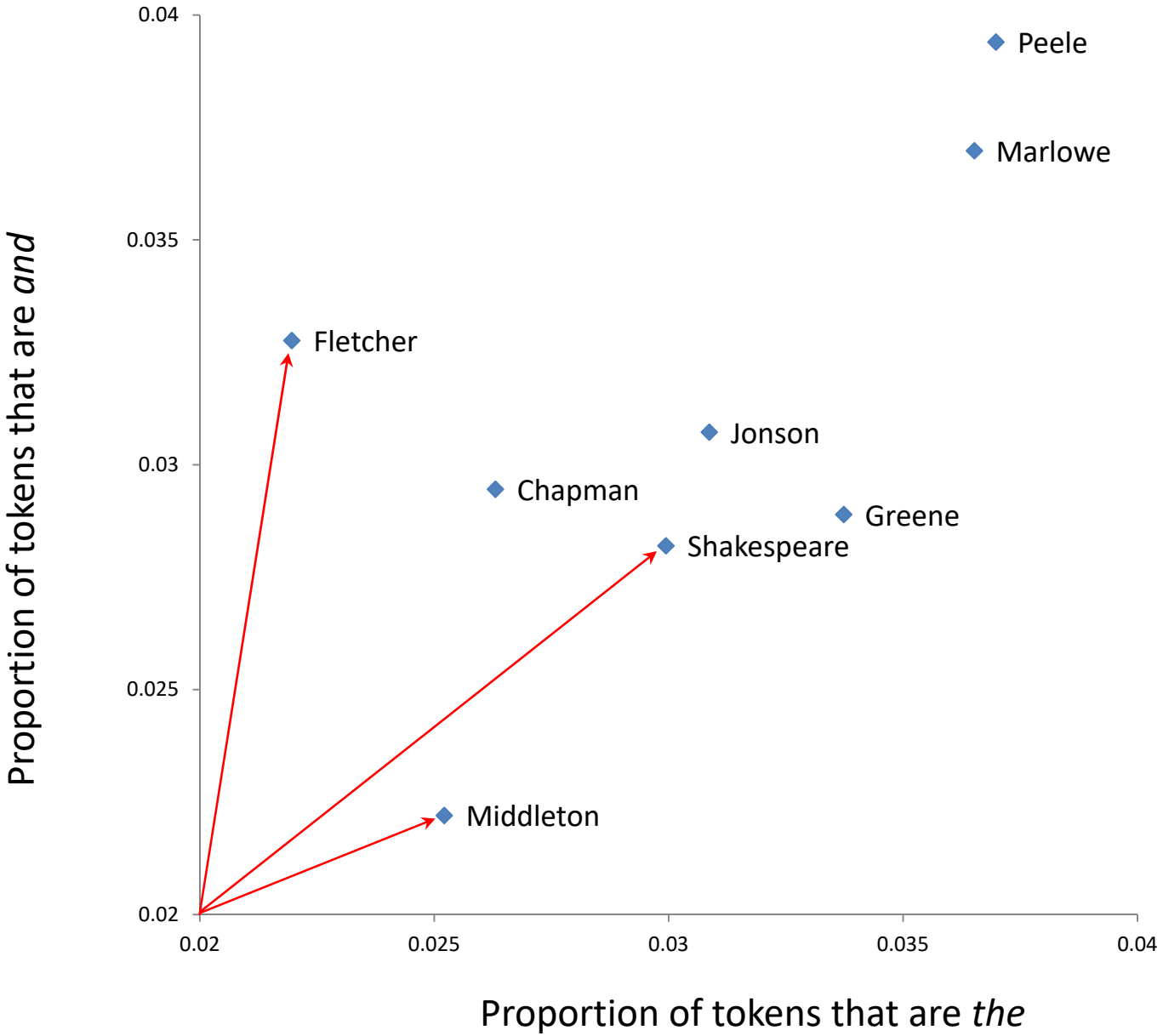
= 28.28

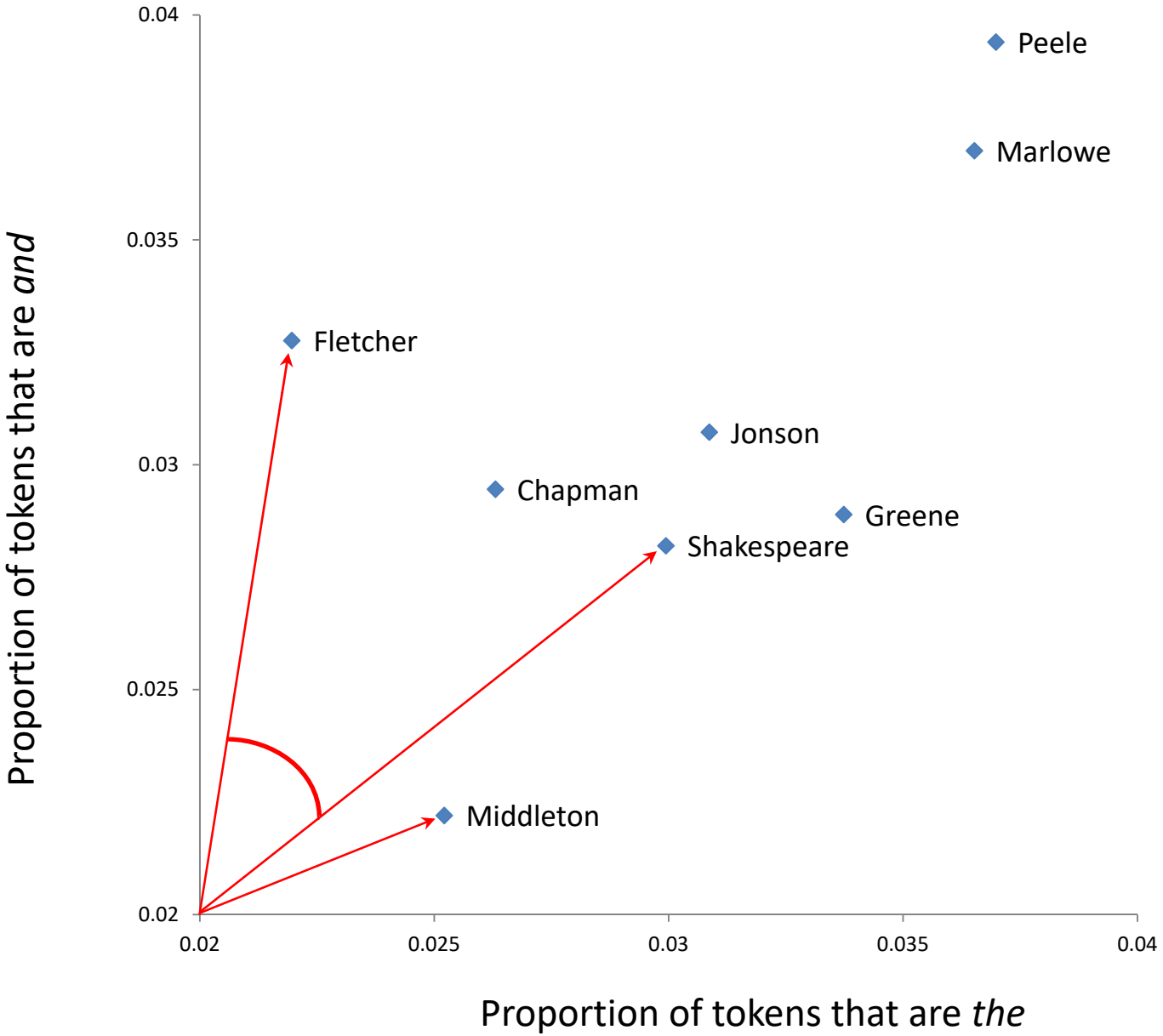
So C is the nearer to A

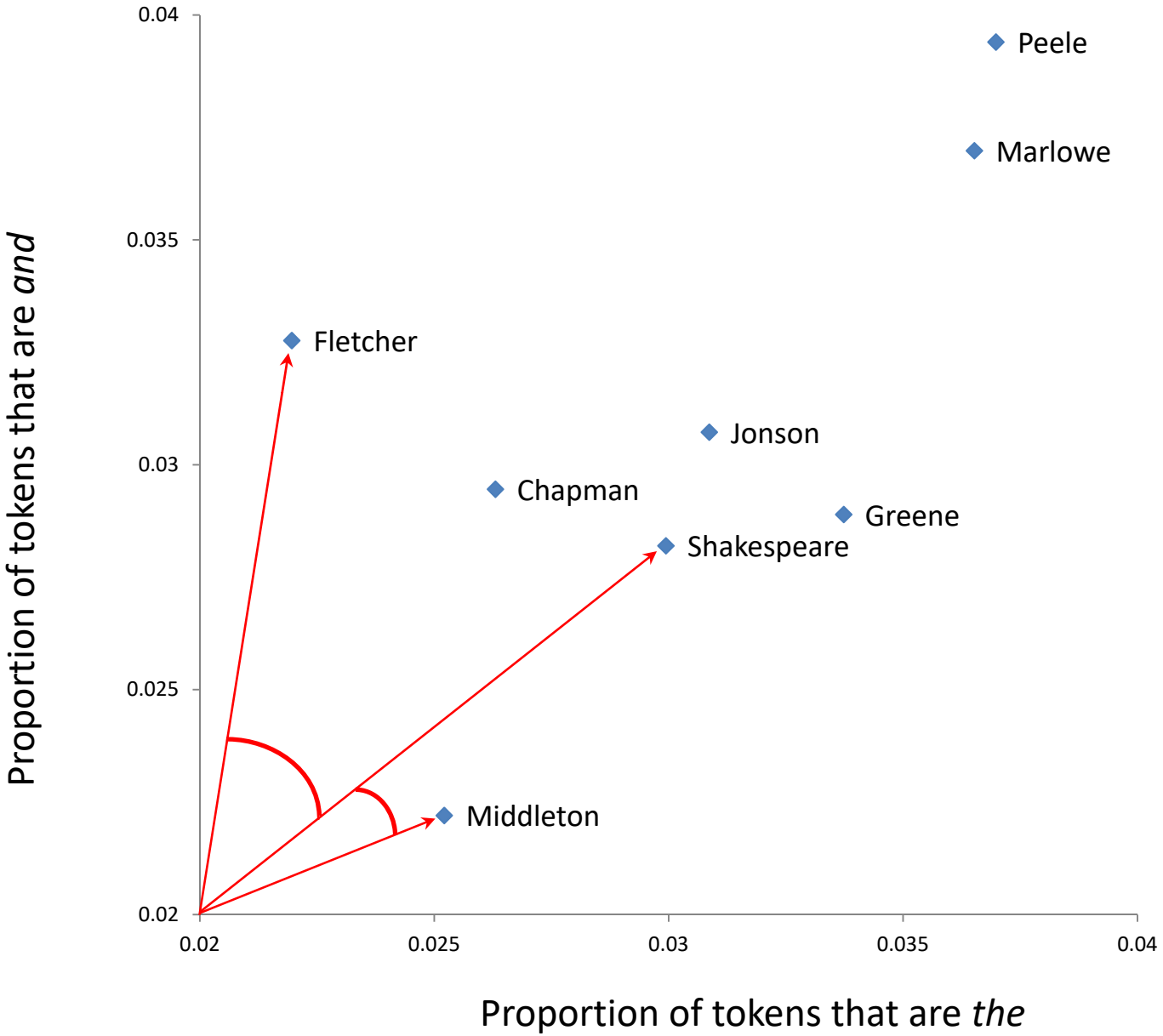

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton
Proportion of tokens that are *and*
Proportion of tokens that are *the*

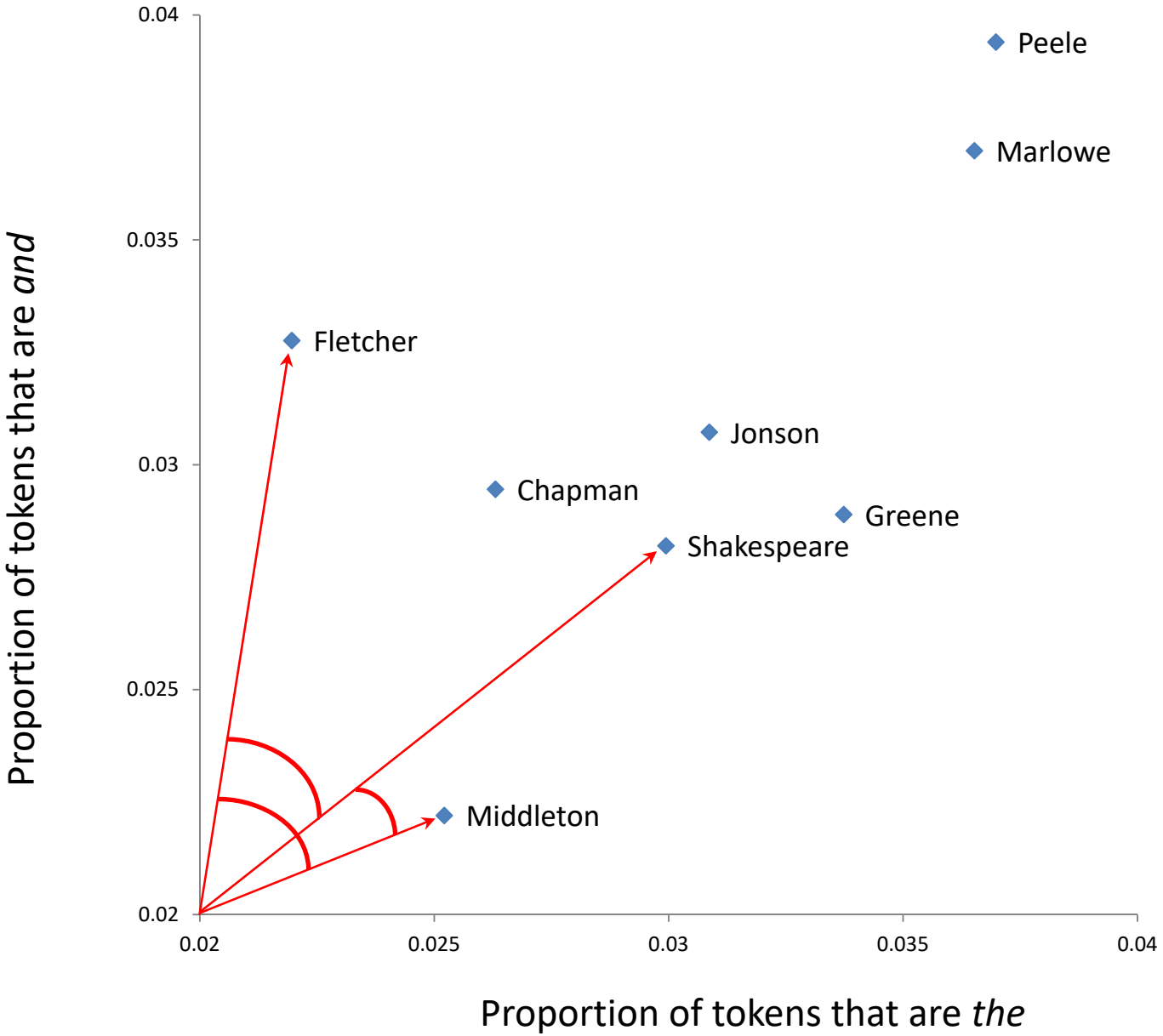

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton
Proportion of tokens that are *and*
Proportion of tokens that are *the*

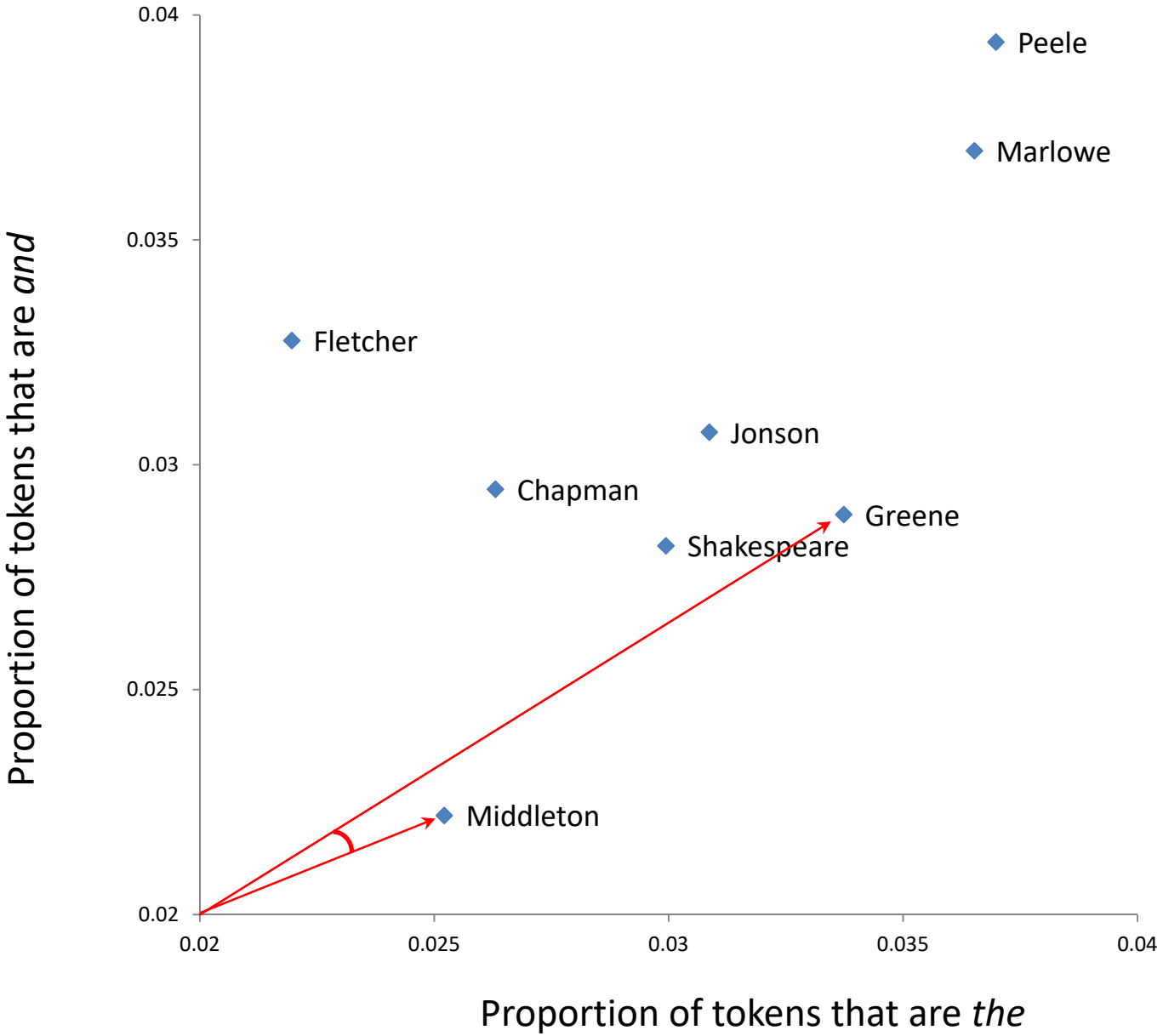

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton
Proportion of tokens that are *and*
Proportion of tokens that are *the*

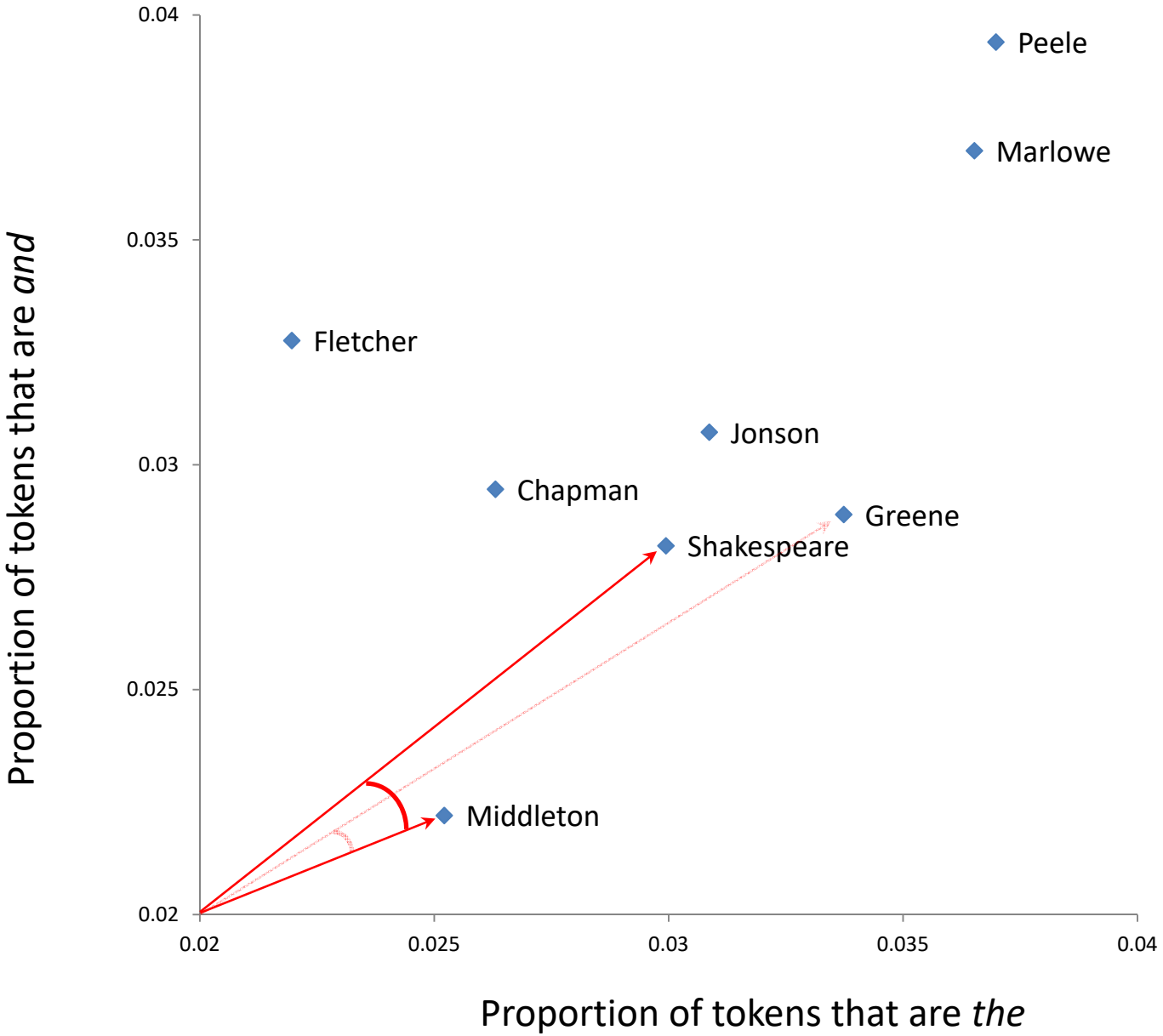

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton
Proportion of tokens that are *and*
Proportion of tokens that are *the*

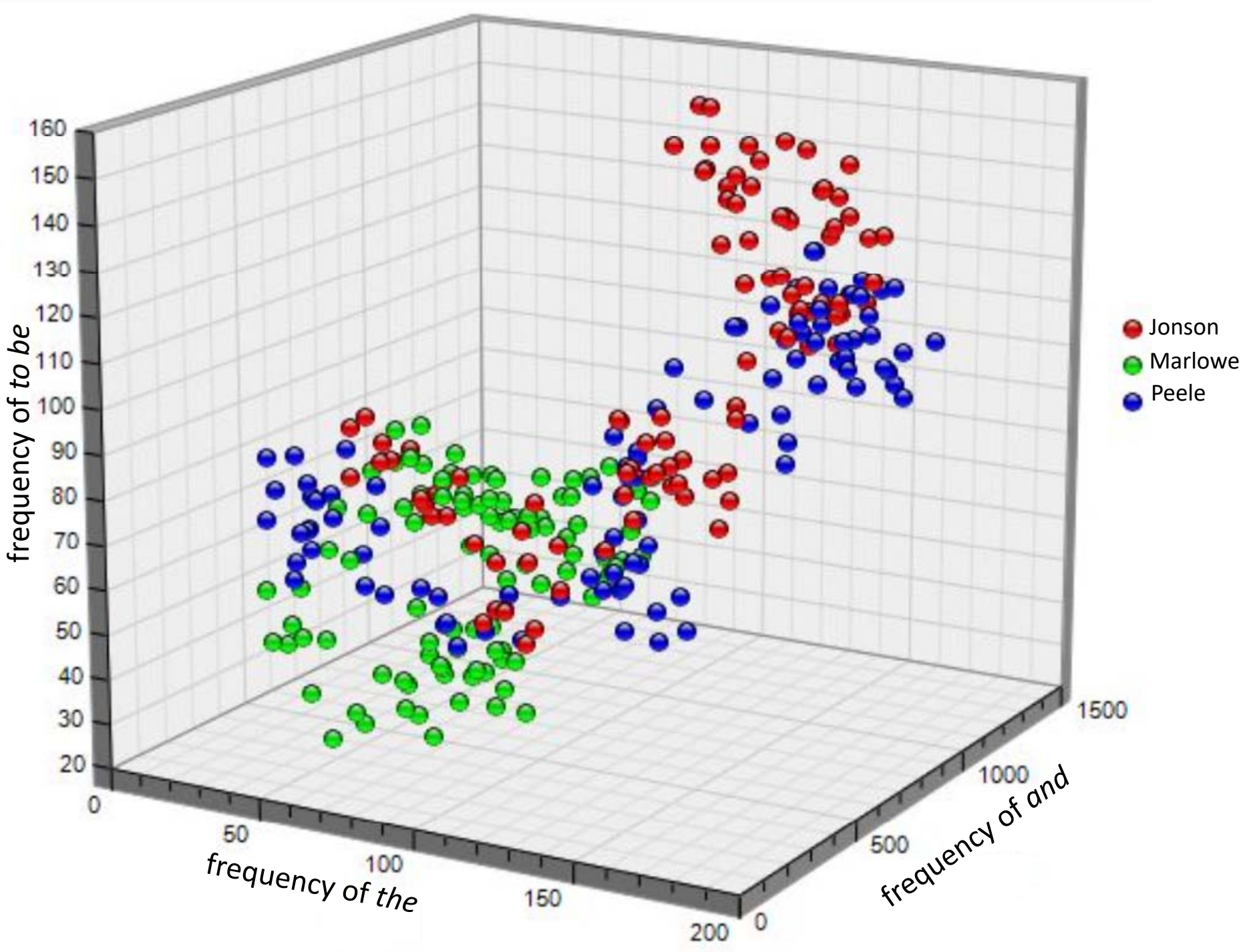

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton
Proportion of tokens that are *and*
Proportion of tokens that are *the*

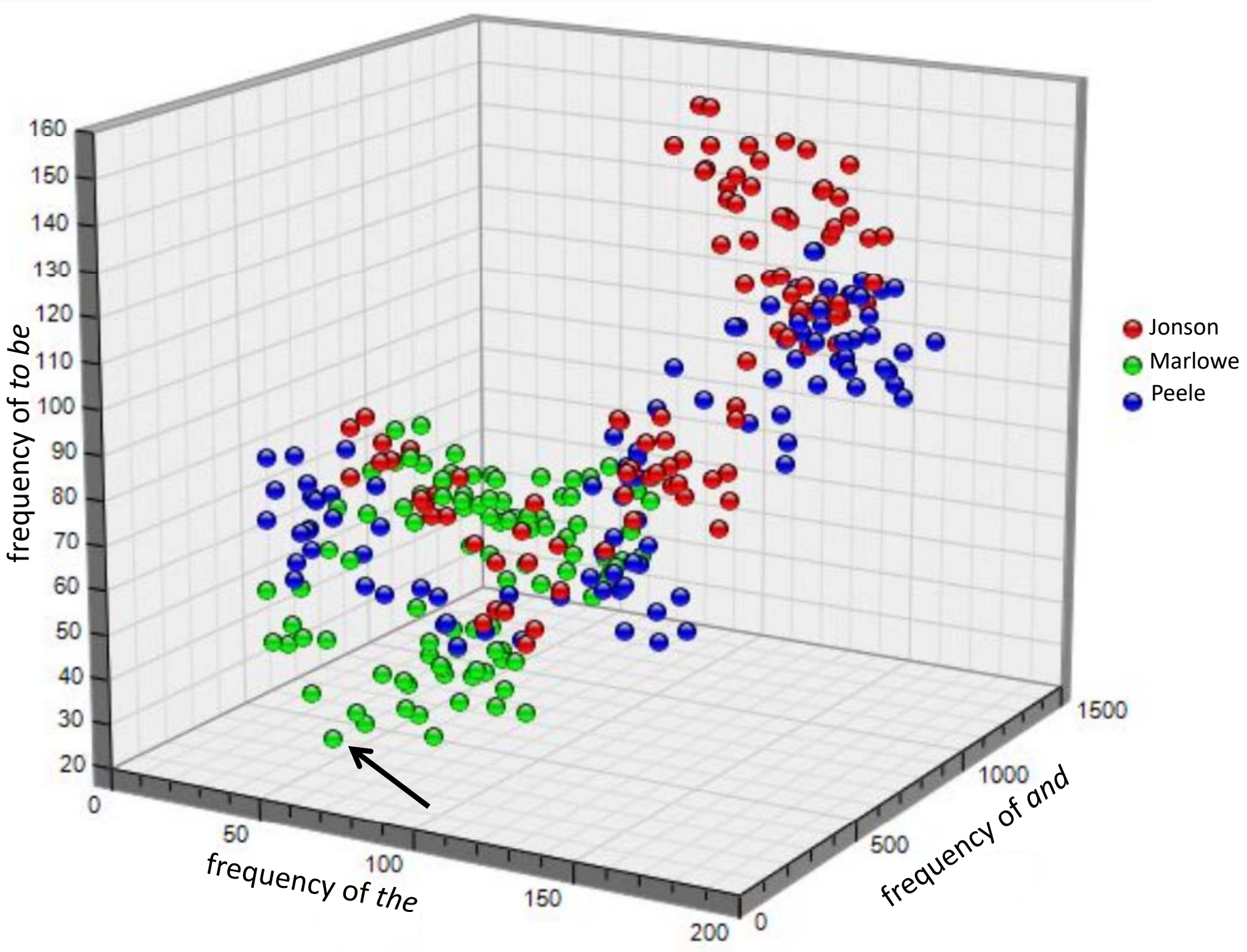

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Proportion of tokens that are *and*
Proportion of tokens that are *the*
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton

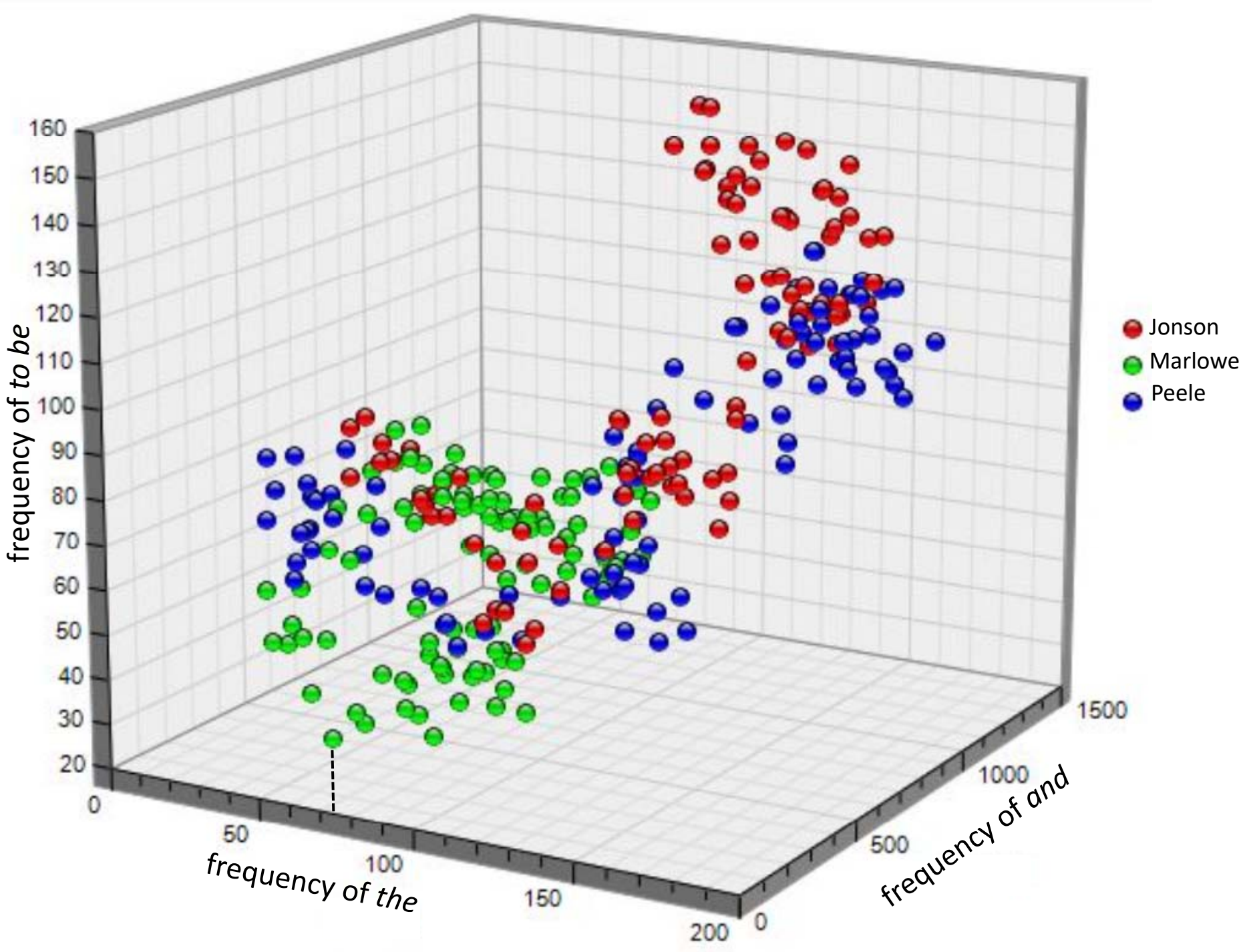

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton
Proportion of tokens that are *and*
Proportion of tokens that are *the*

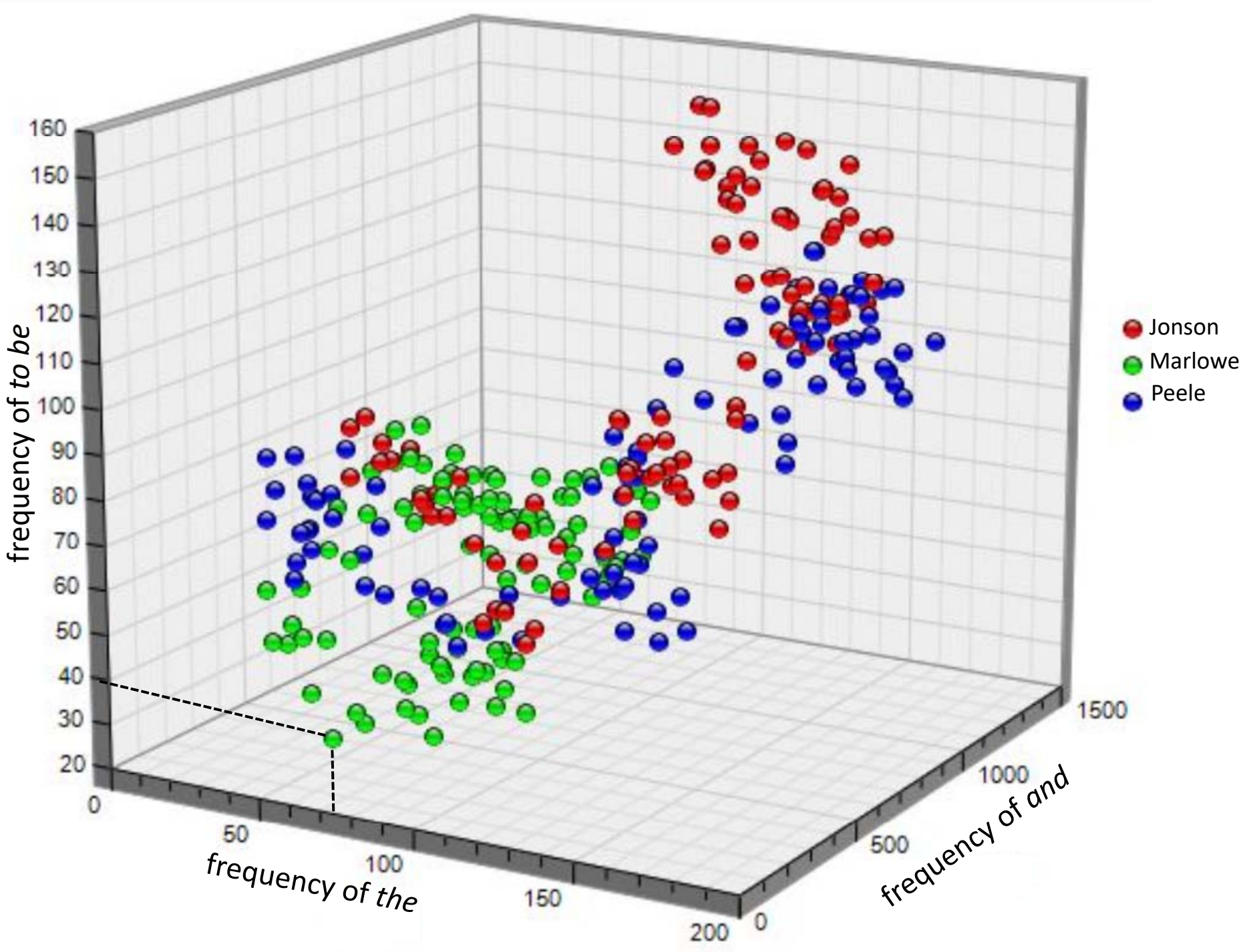

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton
Proportion of tokens that are *and*
Proportion of tokens that are *the*

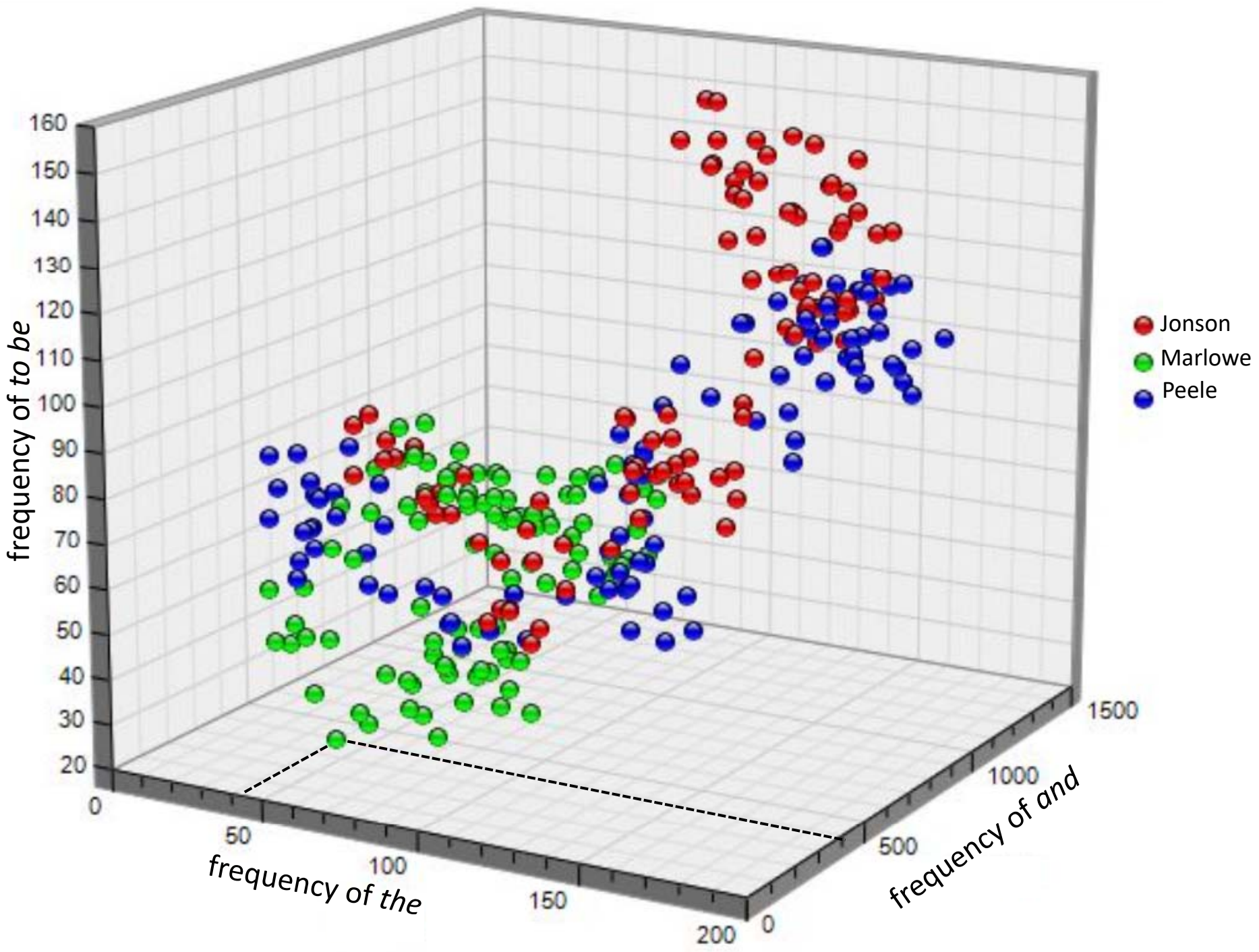

0.04
0.035
0.03
0.025
0.02
0.02
0.025
0.03
0.035
0.04
Peele
Marlowe
Fletcher
Jonson
Chapman
Greene
Shakespeare
Middleton
Proportion of tokens that are *and*
Proportion of tokens that are *the*

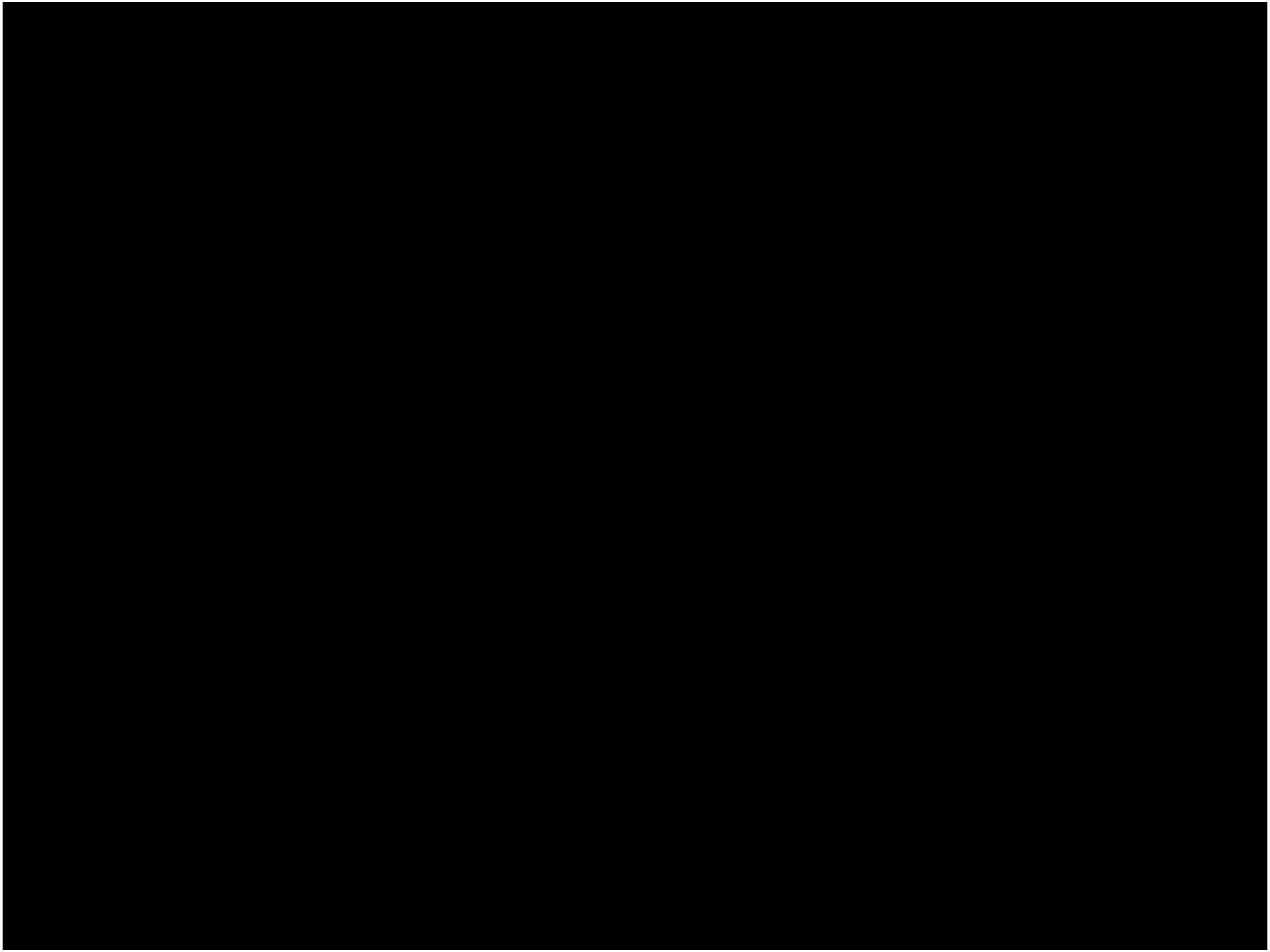

160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
frequency of *to be*
0
50
100
150
200
frequency of *the*
0
500
1000
1500
frequency of *and*
Jonson
Marlowe
Peele

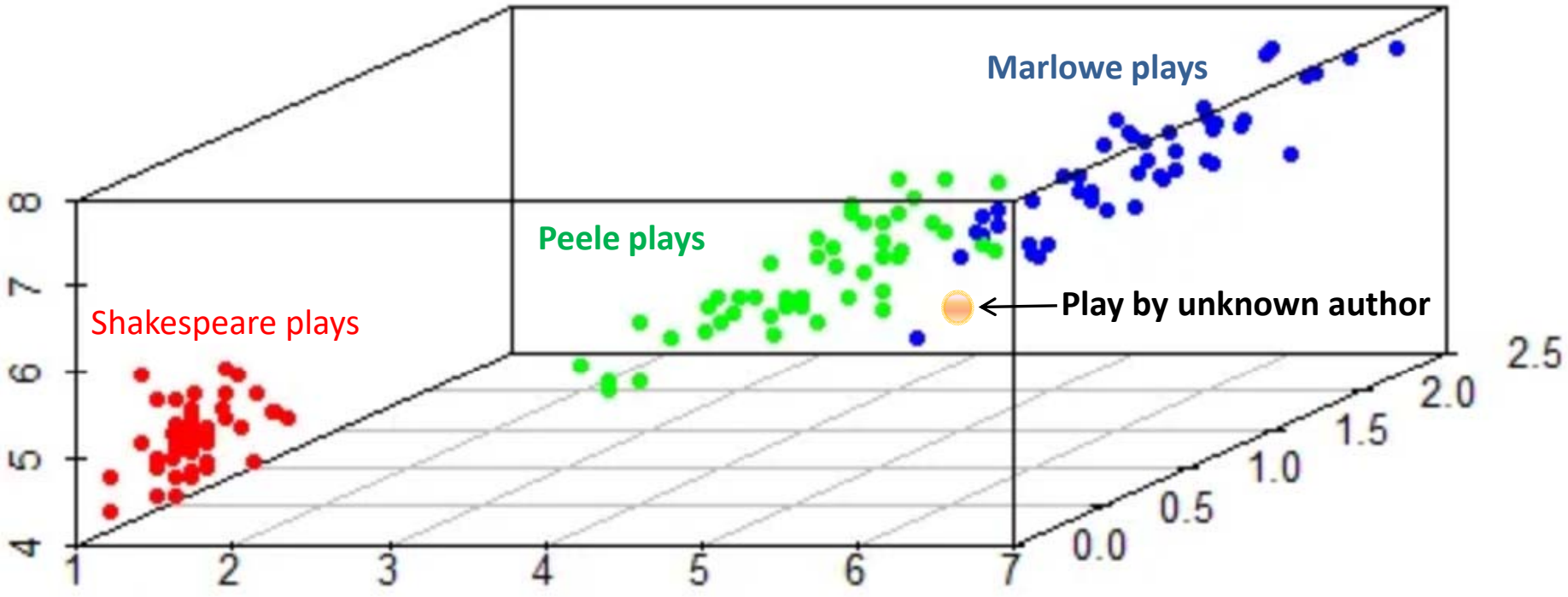

160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
frequency of *to be*
0
50
100
150
200
frequency of *the*
0
500
1000
1500
frequency of *and*
Jonson
Marlowe
Peele


160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
frequency of *to be*
0
50
100
150
200
frequency of *the*
0
500
1000
1500
frequency of *and*
Jonson
Marlowe
Peele

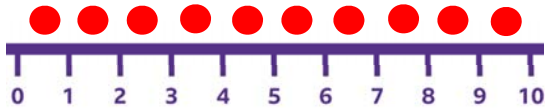

frequency of *to be*
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
0
50
100
150
200
frequency of *the*
0
500
1000
1500
frequency of *and*
Jonson
Marlowe
Peele

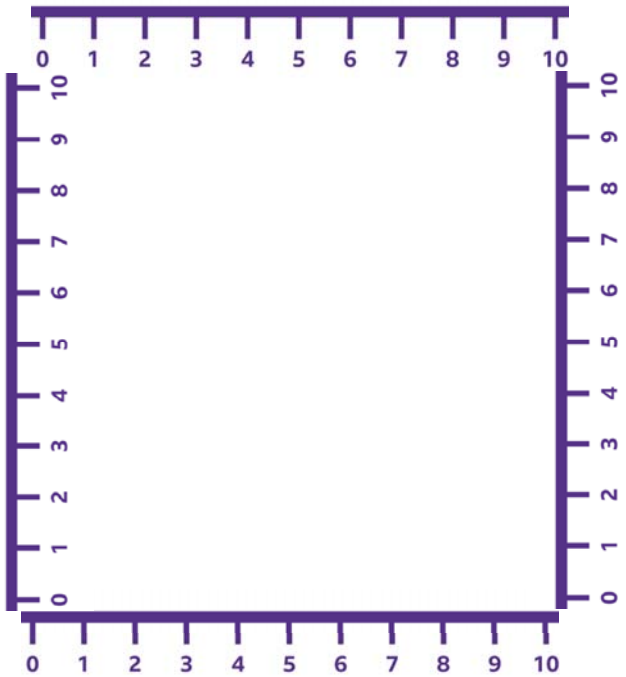

160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
frequency of *to be*
0
50
100
150
200
frequency of *the*
0
500
1000
1500
frequency of *and*
Jonson
Marlowe
Peele

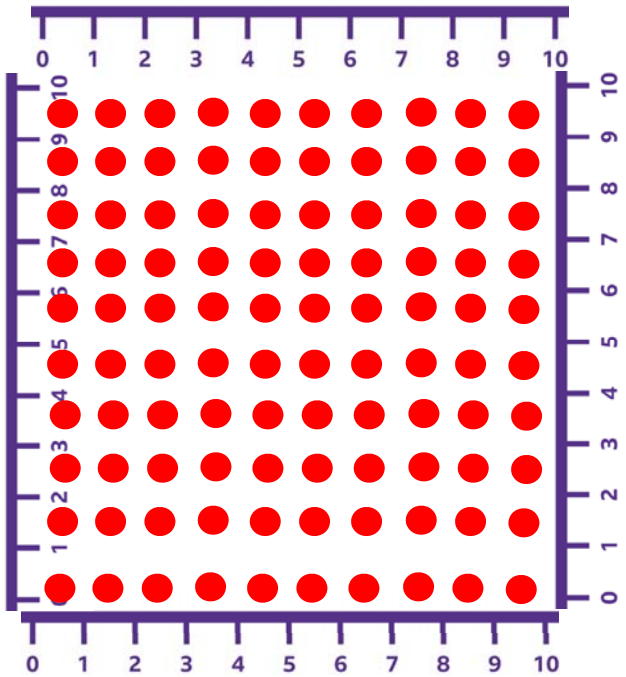


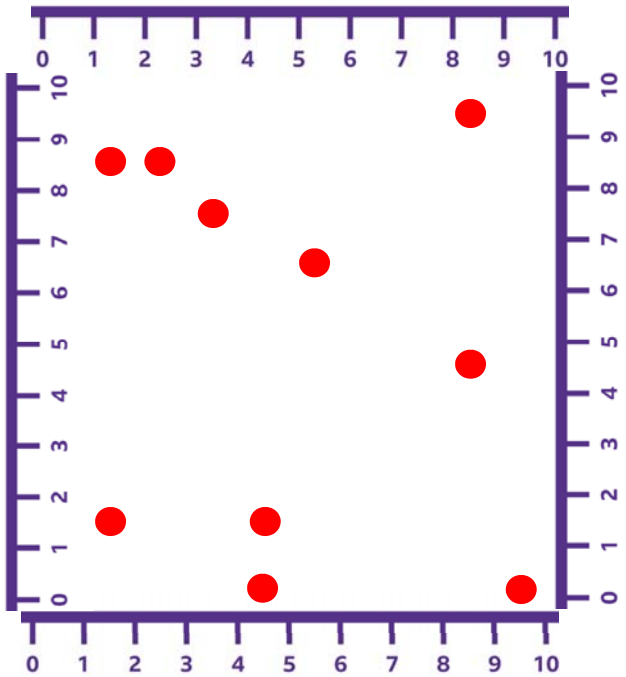

Marlowe plays
Peele plays
Shakespeare plays
Play by unknown author
4
5
6
7
8
1
2
3
4
5
6
7
0.0
0.5
1.0
1.5
2.0
2.5

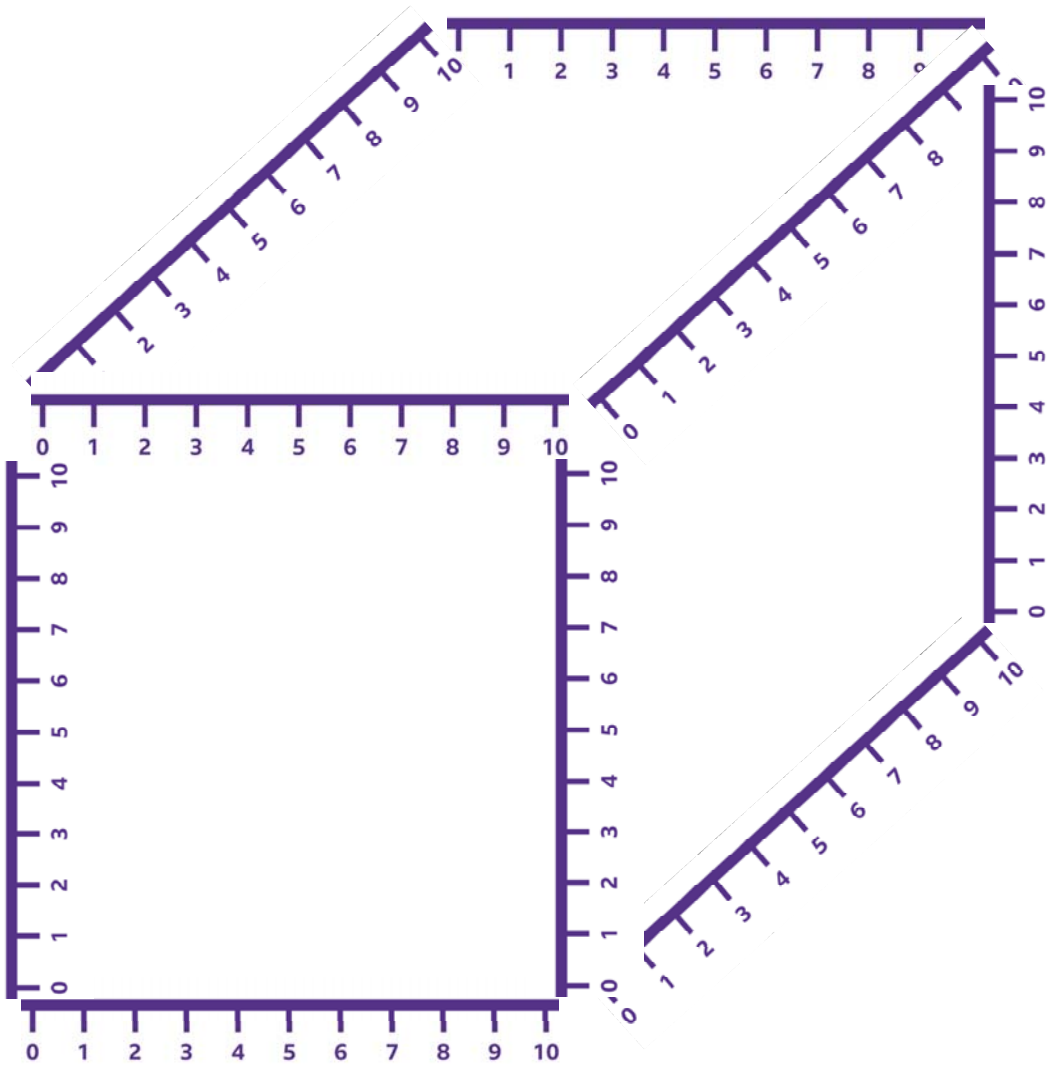
0 1 2 3 4 5 6 7 8 9 10

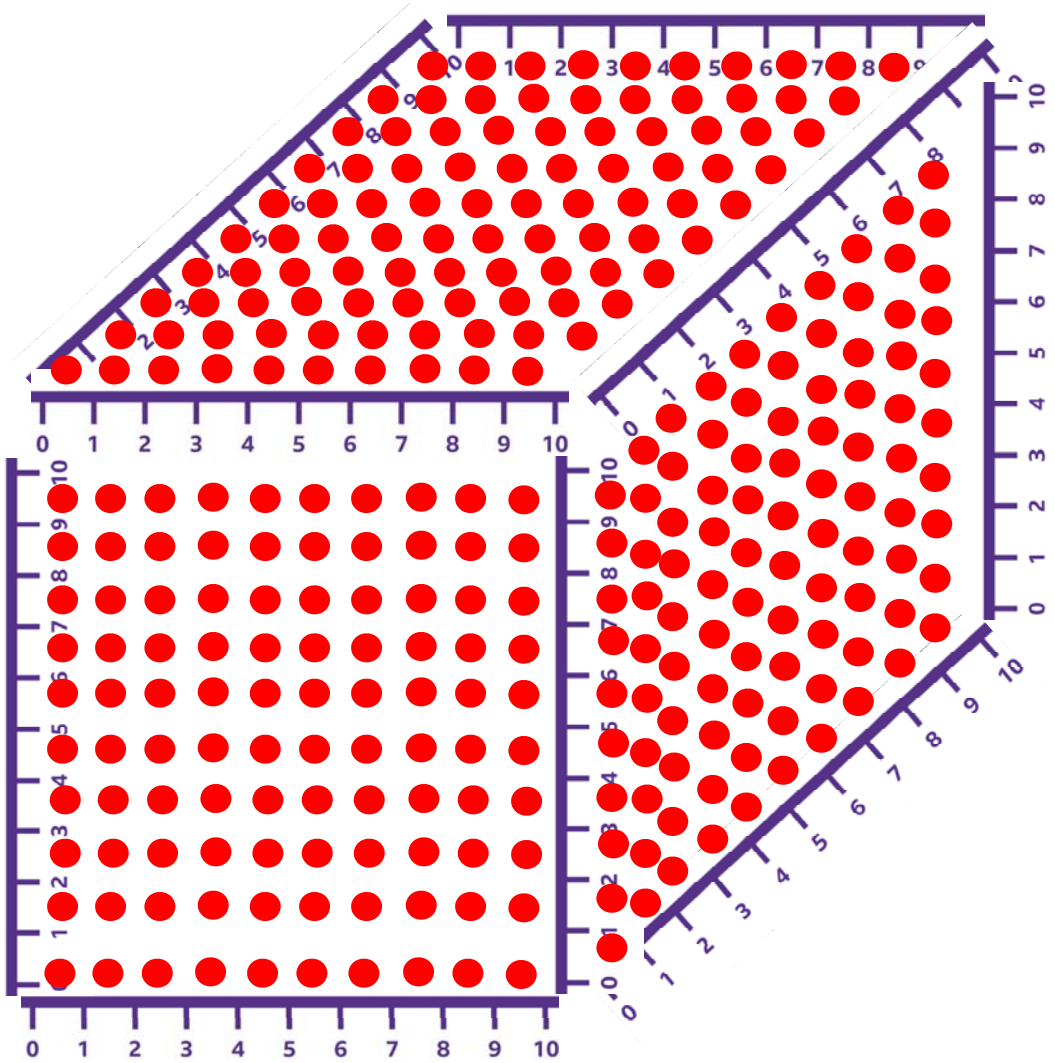

0 1 2 3 4 5 6 7 8 9 10

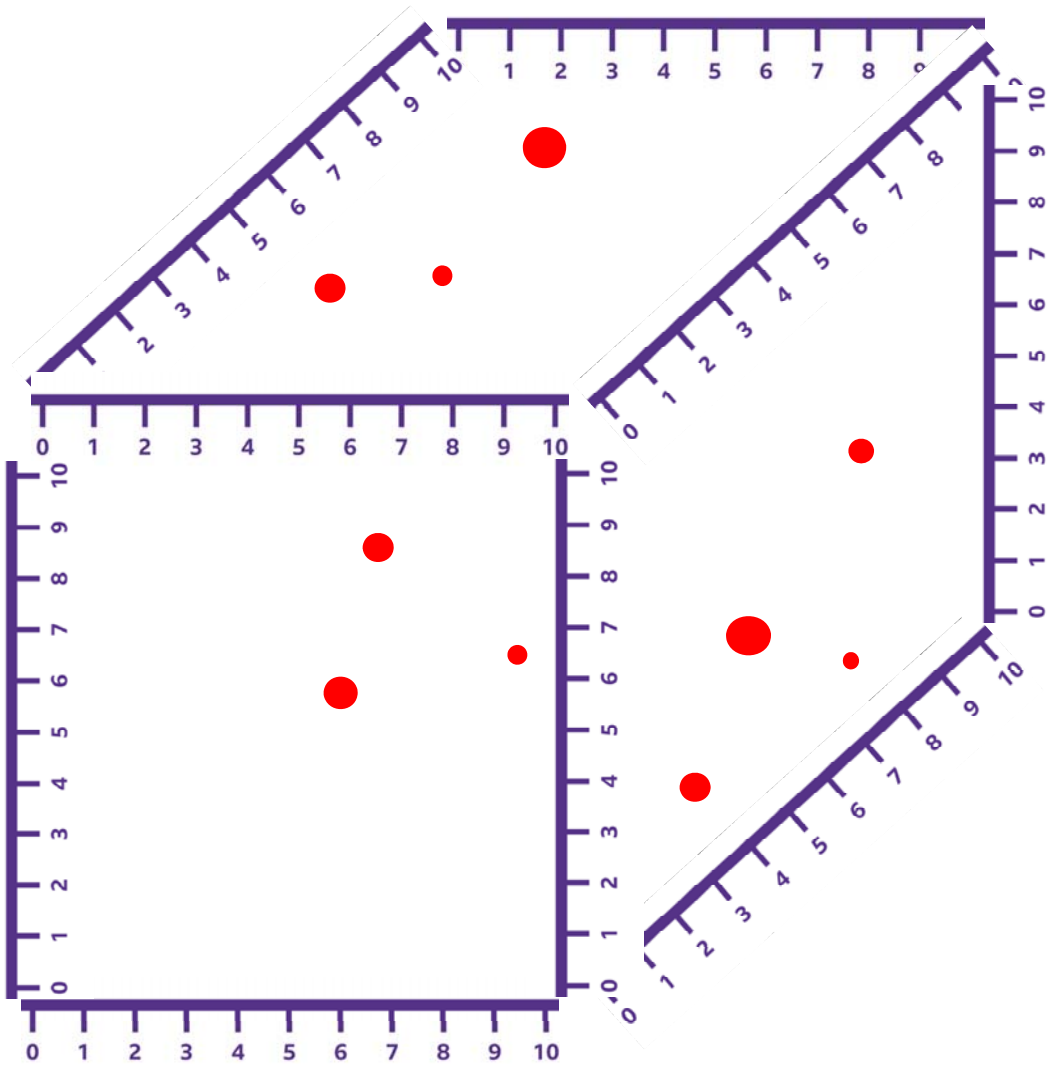

0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5 6 7 8 9 10

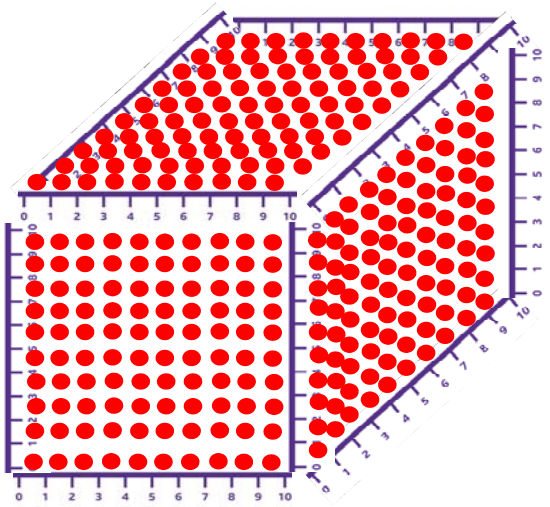

0 1 2 3 4 5 6 7 8 9 10
10 9 8 7 6 5 4 3 2 1 0
0 1 2 3 4 5 6 7 8 9 10


0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5 6 7 8 9 10

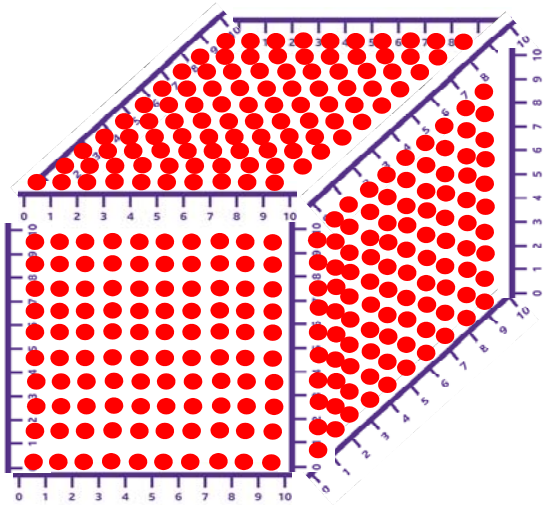

0 1 2 3 4 5 6 7 8 9 10







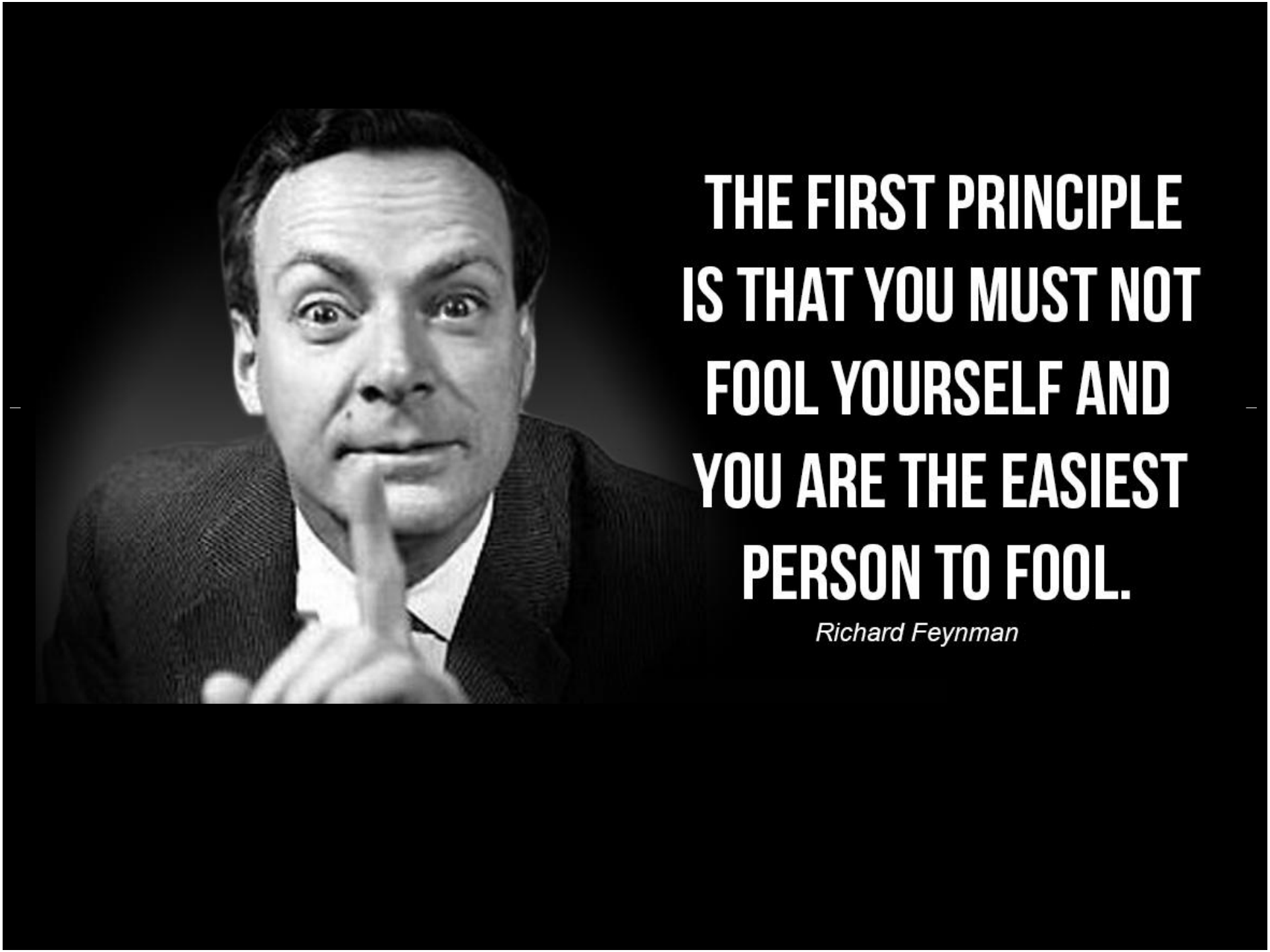
At 100 dimensions we need …



At 100 dimensions we need ...

1,000,000,000,000,000,000,
000,000,000,000,000,000,000,000,000,00
0,000,000,000,000,000,000,000,000,000,
000,000,000,000,000,000,000,000

data points to fill the space.


THE FIRST PRINCIPLE
IS THAT YOU MUST NOT
FOOL YOURSELF AND
YOU ARE THE EASIEST
PERSON TO FOOL.
Richard Feynman